A Coordinate Gradient Descent Method for Structured Nonsmooth Optimization

Sangwoon Yun

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2007

Program Authorized to Offer Degree: Mathematics

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Sangwoon Yun

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

_____

Paul Tseng

Reading Committee:

_____

James Burke

_____

R. Tyrrell Rockafellar

_____

Paul Tseng

Date: _____

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall makes its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Proquest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, 1-800-521-0600, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature_____

Date_____

University of Washington

Abstract

A Coordinate Gradient Descent Method for Structured Nonsmooth Optimization

Sangwoon Yun

Chair of the Supervisory Committee:
Professor Paul Tseng
Department of Mathematics

We consider the problem of minimizing the sum of a smooth function and a (*block separable*) convex function with or without linear constraints. This problem includes as special cases bound-constrained optimization, smooth optimization with ℓ_1-regularization, and linearly constrained smooth optimization such as a large-scale quadratic programming problem arising in the training of support vector machines. We propose a (block) coordinate gradient descent method for solving this class of structured nonsmooth problems. The method is simple, highly parallelizable, and suited for large-scale applications in signal/image denoising, regression, and data mining/classification. We establish global convergence and, under a local Lipschitzian error bound assumption, local linear rate of convergence for this method. The local Lipschitzian error bound holds under assumptions analogous to those for constrained smooth optimization, e.g., the convex function is polyhedral and the smooth function is (nonconvex) quadratic or is the composition of a strongly convex function with a linear mapping. We report our numerical experience with solving the ℓ_1-regularization of unconstrained optimization problems from Moré et al. [73] and from the CUTEr set [38] and some large-scale quadratic program of support vector machines arising from two-class data classification. Comparison with L-BFGS-B and MINOS, applied to a reformulation of the ℓ_1-regularized problem as a bound-constrained smooth op-

timization problem, is reported. Comparison with LIBSVM on large-scale quadratic programming problems of support vector machines is also reported.

In addition, we consider the bi-level problem which minimizes a nonsmooth convex function over the set of stationary points of a certain smooth function. If the smooth function is convex, the convex function is proper, level-bounded, lower semicontinuous, and the set of stationary points of the smooth function over the domain of the convex function is nonempty, a regularization strategy for solving this bi-level problem is proposed for a (block) coordinate gradient descent method. We prove that any cluster point of the generated iterates is a solution of the bi-level problem.

TABLE OF CONTENTS

LIST OF TABLES

GLOSSARY

- QP : quadratic program.
- $\mathcal{N}$: the set of positive integers $1, \ldots, n$.
- x, y, z : vectors in $\mathbb{R}^n$.
- x_j : the jth component of x.
- $x_{\mathcal{J}}$: the subvector of x comprising x_j, $j \in \mathcal{J}$,
- $|\mathcal{J}|$: the cardinality of $\mathcal{J}$
- H, D : $n \times n$ real symmetric matrices.
- $\succ, \succeq$: partial orders relative to the set of positive definite matrices, i.e., $H \succeq D$ (respectively, $H \succ D$) to mean that $H - D$ is positive semidefinite (respectively, positive definite).
- $\lambda_{\min}(H)$, $\lambda_{\max}(H)$: the minimum and maximum eigenvalues of H.
- $H_{\mathcal{J}\mathcal{J}} = (H_{ij})_{i,j \in \mathcal{J}}$: the principal submatrix of H indexed by $\mathcal{J}$.
- I : the identity matrix.
- 0_n : the $n \times n$ matrix of zero entries.
- Null(A) : the null space of the $m \times n$ real matrix A.

- $\|\cdot\|_p$: ℓ_p norm of vectors, i.e., $\|x\|_p = \left(\sum_{j=1}^{n} |x_j|^p\right)^{1/p}$ for $1 \leq p < \infty$.
- $\|\cdot\|$: Euclidean norm of vectors (ℓ_2 norm).
- $\|\cdot\|_\infty$: $\|x\|_\infty = \max_j |x_j|$.

ACKNOWLEDGMENTS

First and foremost I thank my thesis advisor, Professor Paul Tseng, without whom this thesis would never have been completed. His expertise and experience in this field was invaluable. He is always there for advice and has been a constant source of enthusiasm and encouragement throughout the past four years. I appreciate the support I have received while completing this dissertation and in my six years at the University of Washington. I would like to express my sincere thanks to Professors R. Tyrrell Rockafellar and James Burke for their instruction and serving on my supervisory committee. I do not think I could have completed my dissertation without the help of those around me. Thus I truly thank my wife, Jinhee Hong, for her support and dedication to our family and my baby, Paul (Heewoong), for making me smile. Also I thank my parents who paved for me a path to higher education.

Chapter 1
INTRODUCTION

Nonsmooth optimization problems are generally considered to be more difficult than smooth problems. Among those, we study the optimization problems whose objective function is the sum of a smooth function and a structured nonsmooth convex function. In particular, the nonsmooth convex function may be (block) separable or even polyhedral. Such problems arise in bound-constrained optimization, smooth optimization with ℓ_1-regularization, linearly constrained smooth optimization such as large-scale quadratic problems arising in the training of support vector machines and linearly constrained nonsmooth optimization. In applications such as signal/image denoising, regression, and data mining/classification, the problems are highly structured, but large scale. The possibly nonconvex, nonsmooth, and large-scale nature of such problems poses computational challenges. We study a derivative-based method that distributes computation coordinatewise and is highly parallelizable, thus is suitable for large scale problems. Extension to bi-level optimization is also studied.

1.1 A (Block) Coordinate Gradient Descent Method for Nonsmooth Separable Minimization

In Chapter 2, we consider a type of nonconvex nonsmooth optimization problem that arises in many applications and has the following form:

$$\min_x \quad F_c(x) \stackrel{\text{def}}{=} f(x) + cP(x), \tag{1.1}$$

where $c > 0$, $P : \mathbb{R}^n \to (-\infty, \infty]$ is a proper, convex, lower semicontinuous (lsc) function [88], and f is smooth (i.e., continuously differentiable) on an open subset

of $\mathbb{R}^n$ containing $\mathrm{dom}P = \{x \mid P(x) < \infty\}$. Of particular interest is when P has a block-separable structure (see 2.14).

The well studied bound-constrained optimization problem is a special case of (1.1) with

$$P(x) = \begin{cases} 0 & \text{if } l \leq x \leq u; \\ \infty & \text{else,} \end{cases} \tag{1.2}$$

where $l \leq u$ (possibly with $-\infty$ or ∞ components). Such a model arises, e.g., in signal denoising based on Markov random field prior [94]. Another special case of (1.1) that has attracted much interest in signal/image denoising and data mining/classification is when $P(x) = \|x\|_1$. This yields the following problem with ℓ_1-regularization:

$$\min_{x \in \mathbb{R}^n} \quad f(x) + c\|x\|_1. \tag{1.3}$$

For example, f may be the negative of a log-likelihood function. The ℓ_1 term has the desirable property of inducing sparsity in the solution, i.e., few nonzero components, which is useful for finding a sparse representation of a noisy signal or for smoothing a signal/image to have a sparse number of jumps, etc. [1, 7, 14, 20, 21, 68, 92, 93]. In the above cases, P is separable, i.e., a sum of univariate convex lsc functions. Using duality, the "support vector regression" model [7, 23, 107] can be shown to be a special case of (1.1) with P separable and piecewise-linear/quadratic. The group Lasso model for regression [70, 108] is a special case of (1.1) with

$$P(x) = \|x_{\mathcal{J}_1}\|_2 + \cdots + \|x_{\mathcal{J}_N}\|_2, \tag{1.4}$$

where $\mathcal{J}_1, \ldots, \mathcal{J}_N$ is a partition of $\{1, \ldots, n\}$. Here P is block-separable.

This problem (1.1) has previously been studied in [2, 34, 36, 51, 72]. The work most closely related to ours is that of Fukushima and Mine [36], who proposed a proximal gradient descent method which, given $x^k \in \mathrm{dom}P$, computes a direction d^k as the solution of the subproblem

$$\min_d \ \nabla f(x^k)^T d + \frac{1}{2}\rho^k \|d\|^2 + cP(x^k + d)$$

($\rho^k > 0$) and updates $x^{k+1} = x^k + \alpha^k d^k$, with stepsize $\alpha^k > 0$ chosen by the Armijo-type rule. They showed that every cluster point of $\{x^k\}$ is a stationary point of F_c, assuming that ∇f has a Lipschitz continuity property, the directional derivative of P has a continuity property, and ρ^k is uniformly bounded above and below away from zero. Local linear convergence to a stationary point $\bar{x}$ was also shown, assuming that $\nabla^2 f(\bar{x})$ is positive definite. Later, Kiwiel [51] proposed a method in which $P(x^k + d)$ is approximated by a subgradient bundle. Fukushima [34] further extended Kiwiel's method to handle smooth equality constraints via exact penalization, and replaced $\rho^k \|d\|^2$ more generally by a strongly convex proximal term $d^T H^k d$. Mine and Fukushima [72] studied a related Frank-Wolfe-type method corresponding to $\rho^k = 0$, with α^k chosen by line minimization and assuming P is strictly convex. If $\mathrm{dom} P = \mathbb{R}^n$, then (1.1) is a special case of a composite nonsmooth optimization problem studied in [2, 9, 31], i.e., minimizing a real-valued convex function $(t, x) \mapsto t + cP(x)$ composed with a smooth mapping $x \mapsto (f(x), x)$. The descent method of Auslender [2], when specialized to this case, has a form similar to the method of Fukushima and Mine, but with $\rho^k \|d\|^2$ in the objective replaced by a ball constraint $\|d\| \leq 1$ [2, pages 434, 451]. The descent method of Burke [9], when specialized to this case, also has a form similar to the method of Fukushima and Mine, but with $\rho^k \|d\|^2$ replaced more generally by $\rho(d, x^k)$, where ρ belongs to the function class $\mathcal{C}^*$ defined in [9, (3.5)]. Under a certain compactness assumption, every cluster point of $\{x^k\}$ is a stationary point of F_c [2, Theorem 2], [9, Theorem 5.3]. The method of Fletcher [31] uses trust-region instead of line search to achieve global convergence. If in addition f is twice continuously differentiable, then F_c is "lower-C^2" [90, Theorem 10.33], for which locally convergent proximal point methods have been proposed [44, 81, 99]. If f is convex, then an ϵ-subgradient method can also be applied [5, 6, 85]. However, the above studies did not present numerical results, so the practical performance of these methods cannot be judged.

In the special case of bound-constrained smooth optimization, gradient-projection

methods [5, 6, 50, 66, 74] or coordinate descent methods [13, 39, 64, 67, 78] can be effective. Other methods based on trust region or active set, possibly in conjunction with gradient projection to do active-set identification, have also been much studied; see [16, 17, 109] and references therein. In the special case of (1.3), some methods have been proposed for the special case of "basis pursuit," where

$$f(x) = \|Ax - b\|_2^2,$$

the columns of $A \in \mathbb{R}^{m \times n}$ are wavelet functions, and $b \in \mathbb{R}^m$. Specifically, Chen, Donoho and Saunders [14] proposed a primal-dual interior-point (IP) method, with a conjugate-gradient method used to solve the linear equations at each iteration, exploiting the fast multiplications by A and A^T. However, the number of conjugate-gradient steps is large due to ill-conditioning in the linear equations being solved at each IP iteration. For the case where the columns of A comprise the finite union of (overcomplete) sets of orthonormal wavelet packets, Sardy, Bruce, and Tseng [91] proposed an alternative method based on block coordinate descent, which was significantly more efficient than the IP method owing to its fast iterations by exploiting the wavelet structure of A. Although coordinate descent methods do not converge on nonsmooth problems in general, the nonsmooth 1-norm is *separable*, which is key to its convergence. Unfortunately, the coordinate descent method is much less efficient when f is nonquadratic since it requires an expensive coordinate-wise minimization at each iteration; see [40, 41, 92, 93, 94] for further discussions and special cases. Also, if f is nonconvex, then an example of Powell [83] shows that coordinate descent methods can cycle among non-stationary points, even if $P \equiv 0$. Additional assumptions on f are needed to ensure global convergence [101, 102].

We can reformulate (1.1) as a smooth optimization problem over a closed convex set:

$$\min_{x,\xi} \{ f(x) + c\xi \mid P(x) - \xi \leq 0 \}. \tag{1.5}$$

If P is polyhedral, then this problem has linear constraints. The special case of (1.1)

can be reformulated as a bound-constrained smooth optimization problem, though the dimension doubles; see Section 2.6.3. However, although there exist many methods for solving this class of problems (e.g., gradient projection and active-set methods), these methods seem not well suited for the large-scale applications mentioned earlier. In particular, they cannot easily exploit the (block) separable structure of P.

Thus, even in the special case of (1.3), there appears to be no existing method that can efficiently solve this problem when f is nonquadratic and n is large. The nonquadratic case is of practical interest since it allows for non-Gaussian noise in likelihood estimation and includes sparse nonlinear least square problem. We propose a method that can efficiently solve (1.1) and (1.3) on a large scale. Our idea is simple: Since coordinate-wise minimization is expensive when f is nonquadratic, we will replace f in F_c by a strictly convex quadratic approximation. To ensure sufficient descent, we perform an inexact line search on F_c from the current iterate in the direction of the coordinate-wise minimum. Surprisingly, this approach does not appear to have been studied before. Specifically, we propose a (block) coordinate gradient descent (abbreviated as CGD) method for solving (1.1) with P having a block-separable structure. At each iteration, we approximate f by a quadratic and apply block coordinate descent to generate a descent direction. Then we do an inexact line search along this direction and re-iterate. This method is simple, highly parallelizable, and is suited for solving large-scale problems. We show that each cluster point of the iterates generated by this method is a stationary point of F_c, provided that the coordinates are updated in either a Gauss-Seidel manner or a Gauss-Southwell manner; see Theorem 2.1. Thus, coordinate gradient descent not only has cheaper iterations than exact coordinate descent, it also has stronger global convergence properties, able to avoid the aforementioned cycling phenomenon. We next show that if a local Lipschitzian error bound on the distance to the set of stationary points $\bar{X}$ holds and the isocost surfaces of F_c restricted to $\bar{X}$ are properly separated, then the iterates generated by the CGD method converge at least linearly to a stationary point of F_c;

see Theorems 2.2, 2.3 and 2.4. This result is analogous to those obtained for gradient projection, matrix splitting, coordinate descent methods for constrained smooth optimization [60, 61, 62, 63, 100]. We show that this local error bound holds if either (i) f is strongly convex with Lipschitz continuous gradient or (ii) P is polyhedral (not necessarily separable) and f is quadratic or the dual of certain strictly convex function or the composition of a strongly convex function with Lipschitz continuous gradient and an affine mapping; see Theorem 2.5. The proof for case (ii) involves reducing (1.1) to a linearly constrained smooth optimization problem and applying existing error bound results for that problem [25, 60, 61, 62, 84]. In the special case of linearly constrained smooth optimization problem (i.e., P is the indicator function for a polyhedral set), error bound has been much studied and is a key to establishing linear convergence rate for various methods without assuming uniqueness or boundedness of solutions; see [25, 60, 61, 62, 63] and references therein. To our knowledge, error bound for the nonsmooth problem (1.1) has not been studied previously, and the convergence rate analysis involves new proof ideas to handle the nonsmoothness of the objective function F_c. The CGD method may be viewed roughly as a block coordinate version of the method in [36] using a general proximal term, though we also use a different stepsize rule (similar to one in [9]) which is needed for the convergence rate analysis. Our global convergence and convergence rate analyses require weaker assumptions than those in [36].

In Section 2.6, we describe an implementation of the CGD method, along with convergence acceleration techniques, and we report our numerical experience with solving ℓ_1-regularization of nonlinear least square problems from [73] and unconstrained smooth optimization problems from the CUTEr set [38]. We compare the CGD method with L-BFGS-B [109] and MINOS [75], applied to a reformulation of the ℓ_1-regularized problem as a bound-constrained smooth optimization problem. Our comparison suggests that the CGD method can be effective in practice. We discuss conclusions and extensions in Section 2.7.

1.2 A (Block) Coordinate Gradient Descent Method for Linearly Constrained Smooth Optimization and Support Vector Machines Training

In Chapter 3, we consider a linearly constrained smooth optimization problem:

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & f(x) \\ \text{s.t.} & x \in X \stackrel{\text{def}}{=} \{x \mid l \le x \le u,\ Ax = b\}, \end{array} \tag{1.6}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is smooth (i.e., continuously differentiable), $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $l \le u$ (possibly with $-\infty$ or ∞ components). There are many real applications that can be modeled by optimization problems of the form (1.6). For instance, optimal control problems, portfolio selection problems, traffic equilibrium problems, multicommodity network flow problems [3, 18, 46, 71] are specific instances of (1.6). Moreover, an important machine learning methodology, called Support Vector Machine (SVM), leads to huge problems of the form (1.6) with quadratic objective function and $m = 1$.

Support vector machines, invented by Vapnik [106], have been much used for classification and regression, including text categorization, image recognition, hand-written digit recognition, and bioinformatics; see [19] and references therein. The problem of training a SVM may be expressed via duality as a convex quadratic program (QP) with bound constraints plus one equality constraint:

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & \frac{1}{2}x^TQx - e^Tx \\ \text{s.t.} & 0 \le x \le Ce, \\ & a^Tx = 0, \end{array} \tag{1.7}$$

where $a \in \mathbb{R}^n$, $0 < C \le \infty$, $e \in \mathbb{R}^n$ is the vector of all ones, and $Q \in \mathbb{R}^{n\times n}$ is a symmetric positive semidefinite matrix with entries of the form

$$Q_{ij} = a_i a_j K(z_i, z_j), \tag{1.8}$$

with $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ ("kernel function"), and $z_i \in \mathbb{R}^p$ ("ith data point"), $i \in \mathcal{N} \stackrel{\text{def}}{=} \{1, \ldots, n\}$. (Here, "s.t." is short for "subject to".) Popular choices of K are the

linear kernel $K(z_i, z_j) = z_i^T z_j$ (for which $Q = Z^T Z$, with $Z = [a_1 z_1 \ \cdots \ a_n z_n]$, and so rank$Q \leq p$) and the radial basis function (rbf) kernel $K(z_i, z_j) = \exp(-\gamma \|z_i - z_j\|^2)$ where γ is a constant. Often p ("number of features") is not large ($4 \leq p \leq 300$), n is large ($n \geq 5000$), and Q is fully dense and even indefinite; see Section 3.6 for more discussions.

The density and huge size of Q pose computational challenges in solving (1.7). Interior-point methods cannot be directly applied, except in the case of linear kernel where Q has low rank or Q is the sum of a low-rank matrix and a positive multiple of the identity matrix; see [28, 29]. For nonlinear kernel, Fine and Scheinberg [30, Section 4] proposed approximating Q by a low-rank incomplete Cholesky factorization with symmetric permutations. Recently, Scheinberg [96] reported good numerical experience with an active-set method for SVM problems with positive semidefinite Q and, in particular, when the rbf kernel is used. It uses rank-one update of a Cholesky factorization of the reduced Hessian to resolve subproblems. Earlier, Osuna et al. [79] proposed a column-generation approach which solves a sequence of subproblems obtained from (1.7) by fixing some components of x at the bounds. They reported solving problems with up to $n = 100,000$ data points in 200 hours on a Sun Sparc 20. The SVM code in [95] is based on this approach. Motivated by this approach, decomposition methods based on iterative block-coordinate descent were subsequently developed and have become popular for solving (1.7), beginning with the work of Joachims [45], Platt [82], and others, and implemented in SVM codes such as SVM^{light} [45] and LIBSVM [12]. At each iteration of such a method, a small index subset $\mathcal{J} \subseteq \mathcal{N}$ is chosen and the objective function of (1.7) is minimized with respect to the coordinates x_j, $j \in \mathcal{J}$, subject to the constraints and with the other coordinates held fixed at their current value. This minimization needs only those entries of Q indexed by $\mathcal{J}$, which can be quickly generated using (1.8) and, in the case of $|\mathcal{J}| = 2$, has a closed form solution. (We need $|\mathcal{J}| \geq 2$ to satisfy the equality constraint $a^T x = 0$.) Such a method is simple and easy to implement, and for suitable choices of the index

set $\mathcal{J}$, called *working set*, has good convergence properties in theory and in practice. The rows of Q indexed by $\mathcal{J}$ can be cached when updating $\nabla f(x)$ at each iteration, so it need not be recomputed from (1.8) and thus reduces CPU time. Although block-coordinate descent has been well studied for bound constrained optimization (see [6, 67, 103] and references therein), its use for linearly constrained optimization has been little studied prior to SVM.

A good choice of the working set $\mathcal{J}$ is crucial for speed and robustness. In Platt's method [82], which he calls the *sequential minimal optimization* (SMO) method, the working set $\mathcal{J}$ is chosen heuristically with $|\mathcal{J}| = 2$. Joachims [45] proposed the first systematic way of choosing $\mathcal{J}$:

$$\mathcal{J} \in \operatorname*{arg\,min}_{\mathcal{J}':|\mathcal{J}'|\le\ell} \left\{ \begin{array}{ll} \min_d & \nabla f(x)^T d \\ \text{s.t.} & a^T d = 0, \\ & d_j \ge 0, \text{ if } x_j = 0,\ j \in \mathcal{J}', \\ & d_j \le 0, \text{ if } x_j = C,\ j \in \mathcal{J}', \\ & |d_j| \le 1,\ j \in \mathcal{J}', \\ & d_j = 0,\ j \notin \mathcal{J}', \end{array} \right\} \tag{1.9}$$

where $\ell \ge 2$ is an even number. Such $\mathcal{J}$ can be found from among the lowest $\ell/2$ terms from $a_j \nabla f(x)_j$, $j \in \mathcal{I}_+(x) \stackrel{\text{def}}{=} \{j \mid x_j < C, a_j = 1 \text{ or } x_j > 0, a_j = -1\}$ and the highest $\ell/2$ terms from $a_j \nabla f(x)_j$, $j \in \mathcal{I}_-(x) \stackrel{\text{def}}{=} \{j \mid x_j < C, a_j = -1 \text{ or } x_j > 0, a_j = 1\}$, which takes $O(n \min\{\ell, \log n\})$ operations using (partial) sorting. This choice is used in his SVM^{light} code, with $\ell = 10$ as the default value.

Motivated by the aforementioned work, Chang, Hsu and Lin [11] proposed an extension of the SMO method to problems with smooth objective function, in which

the working set is chosen by

$$\mathcal{J} \in \operatorname*{arg\,min}_{\mathcal{J}':|\mathcal{J}'|\leq\ell} \left\{ \begin{array}{ll} \min_d & \nabla f(x)^T d \\ \text{s.t.} & a^T d = 0, \\ & 0 \leq x_j + d_j \leq C,\ j \in \mathcal{J}', \\ & d_j = 0,\ j \notin \mathcal{J}', \end{array} \right\} \tag{1.10}$$

where $\ell \geq 2$. They proved global convergence for their method in that every cluster point of the generated iterates x is a stationary point. Simon [98, Section 6] showed that, in the case of $\ell = 2$, a $\mathcal{J}$ satisfying (1.10) can be found in $O(n)$ operations. For $\ell > 2$, such $\mathcal{J}$ can still be found in $O(n)$ operations [58], though the constant in $O(\cdot)$ depends exponentially in ℓ.

Keerthi et al. [48] proposed choosing, for a fixed tolerance $\epsilon > 0$, a working set $\mathcal{J} = \{i, j\}$ satisfying

$$i \in \mathcal{I}_+(x), \quad j \in \mathcal{I}_-(x), \quad a_i \nabla f(x)_i < a_j \nabla f(x)_j - \epsilon.$$

They proved that the SMO method with this choice of $\mathcal{J}$ terminates in a finite number of iterations with $m(x) \geq M(x) - \epsilon$, where

$$m(x) \stackrel{\text{def}}{=} \min_{j \in \mathcal{I}_+(x)} a_j \nabla f(x)_j, \qquad M(x) \stackrel{\text{def}}{=} \max_{j \in \mathcal{I}_-(x)} a_j \nabla f(x)_j.$$

(Note that a feasible point x of (1.7) is a global minimum if and only if $m(x) \geq M(x)$.) In [49], Keerthi et al. proposed a related choice of $\mathcal{J} = \{i, j\}$ with i and j attaining the minimum and maximum, respectively, in the above definition of $m(x)$ and $M(x)$. This choice, called "maximal violating pair" and used in LIBSVM, is equivalent to Joachim's choice (1.9) with $\ell = 2$.

The first convergence result for the SMO method using the working set (1.9) was given by Lin [53], who proved that every cluster point of the generated iterates x is a global minimum of (1.7), assuming $\min_{\mathcal{J}':|\mathcal{J}'|\leq\ell}(\lambda_{\min}(Q_{\mathcal{J}'\mathcal{J}'})) > 0$. This assumption was later shown by Lin [55] to be unnecessary if $\ell = 2$. Under the further assumptions

that Q is positive definite and strict complementarity holds at the unique global minimum, linear convergence was also proved [54]. List and Simon [57] proposed an extension of the SMO method to problems with more than one linear constraint, in which the working set $\mathcal{J}$ is obtained from maximizing a certain function of x and $\mathcal{J}$. They proved global convergence for their method under the same assumption on Q as Lin. Simon [98] later showed that the maximization subproblem is NP-complete and he proposed a polynomial-time approximation algorithm for finding $\mathcal{J}$ which retains the method's global convergence property.

Hush and Scovel [42] proposed choosing $\mathcal{J}$ to contain a "rate certifying pair", an example of which is (1.10) with $\ell = 2$. They proved that, for any $\epsilon > 0$, the SMO method with this choice of $\mathcal{J}$ terminates in $O(C^2 n^2 (f(x^{\text{init}}) - f(x^*) + n^2\Lambda)/\epsilon)$ iterations with $f(x) \leq f(x^*) + \epsilon$, where x^* is a global minimum of (1.7) and Λ is the maximum norm of the 2×2 principal submatrices of Q. They also showed that a $\mathcal{J}$ satisfying (1.10) can be found in $O(n \log n)$ operations. These complexity bounds were further improved by List and Simon [58] to problems with general linear constraints, where they also showed that a $\mathcal{J}$ satisfying (1.10) can be found in $O(n)$ operations. Hush et al. [43] proposed a more practical choice of $\mathcal{J}$, based on those used in [49] and [98] that achieves the same complexity bounds as in [58].

Palagi and Sciandrone [80] proposed, as a generalization of (1.9), choosing $\mathcal{J}$ to have at most ℓ elements ($\ell \geq 2$) and to contain a maximal violating pair. They also added a proximal term $\tau \|x - x^{\text{current}}\|^2$ to the objective function of (1.7) when minimizing with respect to x_j, $j \in \mathcal{J}$. For this modified SMO method, they proved global convergence with no additional assumption. Chen et al. [15] then proposed a generalization of maximal violating pair by choosing $\mathcal{J} = \{i, j\}$ with $i \in \mathcal{I}_+(x)$, $j \in \mathcal{I}_-(x)$ satisfying

$$a_j \nabla f(x)_j - a_i \nabla f(x)_i \geq \phi(M(x) - m(x)), \tag{1.11}$$

where $\phi : [0, \infty) \to [0, \infty)$ is any strictly increasing function satisfying $\phi(0) = 0$

and $\phi(\alpha) \leq \alpha$ for all $\alpha \geq 0$. Following [80], they also add a proximal term to the objective function, but only when it is not strong convex with respect to x_j, $j \in \mathcal{J}$. For this modified SMO method and allowing Q to be indefinite, they proved global convergence with no additional assumption. Linear convergence was proved for the choice $\phi(\alpha) = \upsilon\alpha$ $(0 < \upsilon \leq 1)$ and under the same assumption as in [54], namely, Q is positive definite and strict complementarity holds at the unique global minimum. While Q can be indefinite for certain kernel functions, the QP (1.7), being nonconvex, can no longer be interpreted as a Lagrangian dual problem.

Fan et al. [26] considered a version of maximal violating pair that uses 2nd-derivative information by adding a Hessian term to the objective of (1.9) with $\ell = 2$:

$$\mathcal{J} \in \operatorname*{arg\,min}_{\mathcal{J}':|\mathcal{J}'|=2} \left\{ \begin{array}{ll} \min_d & \nabla f(x)^T d + \frac{1}{2} d^T Q d \\ \text{s.t.} & a^T d = 0, \\ & d_j \geq 0, \text{ if } x_j = 0,\ j \in \mathcal{J}', \\ & d_j \leq 0, \text{ if } x_j = C,\ j \in \mathcal{J}', \\ & d_j = 0,\ j \notin \mathcal{J}'. \end{array} \right\} \tag{1.12}$$

(This minimizes $f(x + d)$ over all feasible directions d at x with two nonzero components.) However, no fast way for finding such a $\mathcal{J}$ is known beyond checking all $\binom{n}{2}$ subsets of $\mathcal{N}$ of cardinality 2, which is too slow for SVM applications. Fan et al. [26] thus proposed a hybrid strategy of choosing an index i from a maximal violating pair (i.e., $i \in \mathcal{I}_+(x)$ with $a_i \nabla f(x)_i = m(x)$ or $i \in \mathcal{I}_-(x)$ with $a_i \nabla f(x)_i = M(x)$) and then further constraining $\mathcal{J}'$ in (1.12) to contain i. The resulting $\mathcal{J}$ can be found in $O(n)$ operations and improved practical performance. Moreover, such $\mathcal{J}$ belongs to the class of working sets studied in [15], so the convergence results in [15] for a modified SMO method can be applied. Glamachers and Igel [37] proposed a modification of this hybrid strategy whereby if the most recent working set contains an i with $(1 - \delta)C \leq x_i \delta C$ $(0 < \delta < 1/2$, e.g., $\delta = 10^{-8})$, then choose $\mathcal{J}$ by (1.12) with $\mathcal{J}'$ further constrained to contain i; otherwise choose $\mathcal{J}$ to be a maximal violating pair. Glamachers and Igel showed that this choice of $\mathcal{J}$ belongs to the class of working sets

studied in [57], so the convergence result in [57] for the SMO method can be applied. Motivated by this work, Lucidi et al. [59] proposed choosing the working set to be a maximal violating pair $\{i, j\}$ and, if x_i, x_j are strictly between their bounds after the SMO iteration, then performing an auxiliary SMO iteration with respect to a subset $\mathcal{J}'$ of coordinates whose corresponding columns in Q are currently cached. Global convergence for this SMO method was proved under a sufficient descent condition on the auxiliary SMO iteration, which holds if either Q is positive definite or $|\mathcal{J}'| = 2$. Lin et al. [56] proposed a decomposition method for solving the special case of (1.6) with $m = 1$. This method uses a similar line search as our method but generates the descent direction differently, using linear approximations of f instead of quadratic approximations and using working sets $\mathcal{J}$ with $|\mathcal{J}| = 2$ and x_j being "sufficiently free" for some $j \in \mathcal{J}$. Global convergence to stationary points is shown assuming such x_j is uniformly bounded away from its bounds, and improvement over LIBSVM on test problems using the rbf kernel is reported.

We propose a (block) coordinate gradient descent method for solving (1.6) and, in particular, (1.7). At each iteration of our CGD method, a quadratic approximation of f is minimized with respect to a subset of coordinates x_j, $j \in \mathcal{J}$, to generate a feasible descent direction, and an inexact line search on f along this direction is made to update the iterate. For convergence, we propose choosing $\mathcal{J}$ analogously to the Gauss-Southwell-q rule in Chapter 2; see (3.6). We show that each cluster point of the iterates generated by this method is a stationary point of (1.6); see Theorem 3.1. Moreover, if a local error bound on the distance to the set of stationary points $\bar{X}$ of (1.6) holds and the isocost surfaces of f restricted to $\bar{X}$ are properly separated, then the iterates generated by our method converge at least linearly to a stationary point of (1.6); see Theorem 3.2. To our knowledge, this is the first globally convergent block-coordinate update method for general linearly constrained smooth optimization. It has the advantage of simple iterations, and is suited for large scale problems with n large and m small. When specialized to the SVM QP (1.7), our

method is similar to the modified SMO method of Chen et al. [15] and our choice of $\mathcal{J}$ may be viewed as an approximate second-order version of the working set (1.10), whereby a separable quadratic term is added to the objective and $\mathcal{J}$ is chosen as an approximate minimum (i.e., its objective value is within a constant factor of the minimum value). For $m = 1$ and $\ell = 2$, such $\mathcal{J}$ can be found in $O(n)$ operations by solving a continuous quadratic knapsack problem and then finding a conformal realization [89, Section 10B] of the solution; see Section 3.5. Moreover, the local error bound holds for (1.7) always, even if Q is indefinite; see Proposition 3.1. Thus, for SVM, our method is implementable in $O(n)$ operations per iteration and achieves linear convergence without assuming strict complementarity or Q is positive definite as in previous analyses of decomposition methods [15, 26, 54]. We report in Section 3.6 our numerical experience with the CGD method on large SVM QP. Our experience suggests that the method can be competitive with a state-of-the-art SVM code when a nonlinear kernel is used. We give conclusions and discuss extensions in Section 3.7.

1.3 A (Block) Coordinate Gradient Descent Method for Linearly Constrained Nonsmooth Minimization

In Chapter 4, we consider the problem of minimizing the sum of a smooth function and a separable convex function with linear equality constraints:

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & F_c(x) \stackrel{\text{def}}{=} f(x) + cP(x) \\ \text{s.t.} & x \in X \stackrel{\text{def}}{=} \{x \mid l \le x \le u,\ Ax = b\}, \end{array} \tag{1.13}$$

where $c > 0$, f and P are as in Section 1.1, A, b, l, and u are as in Section 1.2. In addition,

$$P(x) = \sum_{j=1}^{n} P_j(x_j),$$

for some proper, convex, lsc functions $P_j : \mathbb{R} \to (-\infty, \infty]$. This problem can be reformulated as an unconstrained problem:

$$\min_x \quad f(x) + c\hat{P}(x), \tag{1.14}$$

with

$$\hat{P}(x) = \begin{cases} P(x) & \text{if } x \in X; \\ \infty & \text{else} \end{cases}.$$

Since $\hat{P}$ is the sum of a convex function and the indicator function of convex set X, $\hat{P}$ is a proper, convex, lsc function. Hence the problem (1.13) is a special case of (1.1) without having a block-separable structure.

Problems of the form (1.13) includes as special cases regularization of smooth optimization problems ($X = \mathbb{R}^n$) and linearly constrained smooth optimization problems ($P \equiv 0$) [34, 36, 51, 72, 103, 104]. Block-coordinate gradient descent methods have been proposed for solving the above two special cases of (1.13) [103, 104] and numerical experiences in [70, 103, 104] suggest that these methods can be effective for solving large problems such as the training of support vector machines ($m = 1$, $P \equiv 0$, f is quadratic) and the ℓ_p-regularization of regression and nonlinear least square problems ($m = 0$, P is the ℓ_1-norm or the sum of ℓ_2-norms). Another special case of (1.13) that arises in the fields such as signal processing is when $f \equiv 0$, $P(x) = \|x\|_1$, and $X = \{x \mid Ax = b\}$ where $Ax = b$ is an underdetermined system of linear equations, in order to obtain a sparse solution [10, 14, 22]. This problem can be reformulated as an LP and be solved efficiently.

In this chapter, we extend the CGD method of Chapters 2 and 3 to solve (1.13). At each iteration of our CGD method, we approximate f by a quadratic and apply block coordinate descent to generate a descent direction. Then we perform an inexact line search along this direction and re-iterate. Following Chapters 2 and 3, we choose the coordinate block according to a Gauss-Southwell-q rule and choose the stepsize according to an Armijo-like rule; see (4.3) and (4.6). (In Chapter 2, a Gauss-Seidel rule and a Gauss-Southwell-r rule for choosing the coordinate block are also considered. For simplicity, we do not consider them here.) Thus, the algorithmic framework is similar to that of Chapters 2 and 3. We give a convergence rate analysis, based on a local Lipschitzian error bound on the distance to the set of stationary points,

that generalizes and unifies those developed in Theorem 2.4 and 3.2 for the two cases of $X = \mathbb{R}^n$ and $P \equiv 0$; see Theorem 4.2. This extension makes use of a new lemma (see Lemma 4.5) and does not assume P is block-separable or $P \equiv 0$. We give the first complexity analysis for the CGD method in the case where f is convex with Lipschitz continuous gradient; see Theorem 4.3. When specialized to the training of support vector machines ($m = 1$, $P \equiv 0$, f is quadratic), our overall complexity bound of $O\left(\frac{n^3 \Lambda b_{\max}^2}{\epsilon} + n^2 \Lambda \max\left\{0, \ln\left(\frac{(F_c(x^{\text{init}}) - \min_{x \in X} F_c(x))}{n b_{\max}}\right)\right\}\right)$ operations, where $b_{\max} = \max_{1 \le i \le n}(u_i - l_i)$ and Λ is the maximum norm of the 2×2 principal submatrices of $\nabla^2 f(x)$, for achieving ϵ-optimality compares favorably with existing bounds of [42, 58]; see Section 4.5 for more details. We show that, in the case where P is separable and polyhedral, the Gauss-Southwell-q rule is implementable in linear-time when $m = 1$ and in polynomial-time when $m > 1$; see Section 4.5. This extends the procedure in Section 3.5, based on finding a conformal realization [89, Section 10B] of a vector in Null(A), for the case of $P \equiv 0$.

1.4 A (Block) Coordinate Gradient Descent Method for Bi-level Optimization

In Chapter 5, we consider the bi-level problem:

$$\min_{x \in S_f} P(x), \tag{1.15}$$

where P is as in Section 1.1 and S_f denotes the set of stationary points of a smooth function f over $\text{dom}P = \{x \mid P(x) < \infty\}$, which we assume to be closed.

The well-known linear least square problem:

$$\min_x \quad \|Ax - b\|_2^2, \tag{1.16}$$

where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, is a challenging computational problem arising in structural engineering, numerical geodesy, and numerical optimization, etc. In many applications, components of x are parameters that must lie within certain bounds.

This leads to a convex quadratic programming problem with bound constraints:

$$\min_{l \le x \le u} \quad \|Ax - b\|_2^2,$$

where $l \le u$ (possibly with $-\infty$ or ∞ components).

When $n > m$, the solution of the least square problem (1.16) may be far from unique. Recent interests have focused on finding solutions that are sparse, i.e., have few nonzero components. In the "Basis Pursuit" model for signal denoising [14, 20, 21], an ℓ_1-regularization term is added to the linear least square:

$$\min_x \quad \|Ax - b\|_2^2 + c\|x\|_1, \tag{1.17}$$

where $c > 0$ is a user chosen regularization parameter. Another way to find an sparse solution is to solve the following bi-level problem:

$$\min_x \qquad \|x\|_1 \qquad \text{where } x \text{ solves (1.16)}. \tag{1.18}$$

If the linear equation $Ax = b$ has at least one solution, then (1.18) is equivalent to the following problem:

$$\min_x \qquad \|x\|_1 \qquad \text{subject to } Ax = b. \tag{1.19}$$

This problem can be reformulated as an LP and be solved efficiently. The problem (1.18) is a special case of the bi-level problem (1.15).

For a primal-dual IP method and a (block) SOR method, the strategy that dynamically adjusts the regularization parameter c of (1.17) towards 0 so that the generated points approach a solution of the bi-level problem (1.18) was proposed with c decreasing at a rate depending on a certain measure of current solution accuracy [93]. More significantly, the convergence rate was not adversely affected as $c \to 0$, i.e., the problem did not become more ill-conditioned with small c.

In this chapter, for the CGD method, we propose a regularization strategy which decreases the regularization parameter c of (1.1) to 0 to solve (1.15). For fixed c,

we solve (1.1) by using the CGD method and then if the generated points reach the desired threshold, we decrease c. At each iteration k ($k = 1, 2, \ldots$), a regularization parameter $c^k > 0$ and an accuracy tolerance ϵ^k are chosen, and the CGD method is applied to (1.1) with $c = c^k$ until it finds an approximate solution x^k satisfying suitable measures of solution accuracy; see Section 5.3.

We show that if f is convex, P is a level-bounded function, and $S_f \neq \emptyset$, then any cluster point of the generated iterates is a solution of the bi-level problem (1.15); see Theorem 5.1.

Chapter 2

A (BLOCK) COORDINATE GRADIENT DESCENT METHOD FOR NONSMOOTH SEPARABLE MINIMIZATION

In this chapter, we study the CGD method for solving (1.1), in particular, (1.3). We describe the CGD method formally and show the global convergence and asymptotic convergence rate of the method when the coordinates are updated in either a Gauss-Seidel manner or a Gauss-Southwell manner. We compare the CGD method with L-BFGS-B and MINOS, applied to solving ℓ_1-regularization of large nonlinear least square problems from Moré et al. [73]. Our comparison suggests that the CGD method is more robust than L-BFGS-B and is faster than MINOS. This chapter is based on the paper [103] co-authored with P. Tseng.

2.1 (Block) Coordinate Gradient Descent Method

In our method, we use $\nabla f(x)$ to build a quadratic approximation of f at x and apply coordinate descent to generate an improving direction d at x. More precisely, we choose a nonempty index subset $\mathcal{J} \subseteq \mathcal{N}$ and a symmetric matrix $H \succ 0_n$ (approximating the Hessian $\nabla^2 f(x)$), and move x along the direction $d = d_H(x; \mathcal{J})$, where

$$d_H(x;\mathcal{J}) \stackrel{\text{def}}{=} \arg\min_{d\in\mathbb{R}^n} \left\{ \nabla f(x)^T d + \frac{1}{2} d^T H d + cP(x+d) - cP(x) \mid d_j = 0 \ \forall j \notin \mathcal{J} \right\}. \tag{2.1}$$

Notice that $d_H(x;\mathcal{J})$ depends on H only through $H_{\mathcal{J}\mathcal{J}}$. This coordinate gradient descent approach may be viewed as a hybrid of gradient-projection and coordinate descent, with connection to the variable/gradient distribution methods for uncon-

strained smooth optimization [27, 35, 65]. In particular,

- if $\mathcal{J} = \mathcal{N}$ and P is given by (1.2), then d is a scaled gradient-projection direction for bound-constrained minimization [6, 50, 74, 77];

- if f is quadratic and we choose $H = \nabla^2 f(x)$, then d is a (block) coordinate descent direction [6, 77, 91, 101, 102].

If H is block-diagonal and P is accordingly block-separable, then (2.1) decomposes into subproblems that can be solved in parallel.

Using the convexity of P, we have the following lemma showing that d is a descent direction at x whenever $d \neq 0$.

Lemma 2.1 *For any $x \in \mathrm{dom} P$, nonempty $\mathcal{J} \subseteq \mathcal{N}$ and $H \succ 0_n$, let $d = d_H(x; \mathcal{J})$ and $g = \nabla f(x)$. Then*

$$F_c(x + \alpha d) \leq F_c(x) + \alpha \left(g^T d + cP(x + d) - cP(x) \right) + o(\alpha) \quad \forall \alpha \in (0, 1], \tag{2.2}$$

$$g^T d + cP(x + d) - cP(x) \leq -d^T H d. \tag{2.3}$$

Proof. For any $\alpha \in (0, 1]$, we have from the convexity of P and $H \succ 0_n$ that

$$\begin{aligned} F_c(x + \alpha d) - F_c(x) &= f(x + \alpha d) - f(x) + cP(\alpha(x + d) + (1 - \alpha)x) - cP(x) \\ &\leq f(x + \alpha d) - f(x) + \alpha cP(x + d) + (1 - \alpha)cP(x) - cP(x) \\ &= \alpha g^T d + o(\alpha) + \alpha(cP(x + d) - cP(x)), \end{aligned}$$

which proves (2.2).

For any $\alpha \in (0, 1)$, we have from (2.1) and the convexity of P that

$$\begin{aligned} & g^T d + \frac{1}{2} d^T H d + cP(x + d) - cP(x) \\ \leq\ & g^T(\alpha d) + \frac{1}{2}(\alpha d)^T H(\alpha d) + cP(x + \alpha d) - cP(x) \\ \leq\ & \alpha g^T d + \frac{1}{2}\alpha^2 d^T H d + \alpha cP(x + d) - \alpha cP(x). \end{aligned}$$

Rearranging terms yields

$$(1-\alpha)g^T d + (1-\alpha)(cP(x+d) - cP(x)) + \frac{1}{2}(1-\alpha^2)d^T H d \leq 0.$$

Since $1-\alpha^2 = (1-\alpha)(1+\alpha)$, dividing both sides by $1-\alpha > 0$ and then taking $\alpha \uparrow 1$ prove (2.3). ■

The bound (2.3) is sharp when $P \equiv 0$. We next choose a stepsize $\alpha > 0$ so that $x' = x + \alpha d$ achieves sufficient descent, and re-iterate. We now describe formally the block coordinate gradient descent method.

CGD method:

Choose $x^0 \in \mathrm{dom} P$. For $k = 0, 1, 2, ...$, generate x^{k+1} from x^k according to the iteration:

1. Choose a nonempty $\mathcal{J}^k \subseteq \mathcal{N}$ and an $H^k \succ 0_n$.

2. Solve (2.1) with $x = x^k,\ \mathcal{J} = \mathcal{J}^k,\ H = H^k$ to obtain $d^k = d_{H^k}(x^k; \mathcal{J}^k)$.

3. Choose a stepsize $\alpha^k > 0$ and set $x^{k+1} = x^k + \alpha^k d^k$.

Various stepsize rules for smooth optimization [6, 32, 33, 77] can be extended to our nonsmooth setting to choose α^k. The following adaptation of the Armijo rule, based on Lemma 2.1 and [9, Subsections 4.2, 4.3], is simple and seems effective from both theoretical and practical standpoints.

Armijo rule:

Choose $\alpha^k_{\mathrm{init}} > 0$ and let α^k be the largest element of $\{\alpha^k_{\mathrm{init}}\beta^j\}_{j=0,1,...}$ satisfying

$$F_c(x^k + \alpha^k d^k) \leq F_c(x^k) + \alpha^k \sigma \Delta^k, \tag{2.4}$$

where $0 < \beta < 1$, $0 < \sigma < 1$, $0 \leq \theta < 1$, and

$$\Delta^k \stackrel{\mathrm{def}}{=} \nabla f(x^k)^T d^k + \theta {d^k}^T H^k d^k + cP(x^k + d^k) - cP(x^k). \tag{2.5}$$

Since $H^k \succ 0_n$ and $0 \leq \theta < 1$, we see from Lemma 2.1 that

$$F_c(x^k + \alpha d^k) \leq F_c(x^k) + \alpha \Delta^k + o(\alpha) \quad \forall \alpha \in (0, 1],$$

and $\Delta^k \leq (\theta - 1) d^{k^T} H^k d^k < 0$ whenever $d^k \neq 0$. Since $0 < \sigma < 1$, this shows that α^k given by the Armijo rule is well defined and positive. This rule, like that for sequential quadratic programming methods [6, 9, 17, 32, 34, 77], requires only function evaluations. And, by choosing α^k_{init} based on the previous stepsize α^{k-1}, the number of function evaluations can be kept small in practice. Notice that Δ^k increases with θ. Thus, larger stepsizes will be accepted if we choose either σ near 0 or θ near 1. The descent condition (2.4) is similar to those used in [2, 9] and the term Δ^k therein seems essential to our convergence rate analysis; see Section 2.7 for discussions.

For convergence, the index subset $\mathcal{J}^k$ must be chosen judiciously. For smooth optimization, $\mathcal{J}^k$ is often chosen in a *Gauss-Seidel* manner, e.g., $\mathcal{J}^k$ cycles through $\{1\}, \{2\}, ..., \{n\}$ or, more generally, $\mathcal{J}^0, \mathcal{J}^1, ...$ collectively covers $1, 2, ..., n$ for every T consecutive iterations, where $T \geq 1$ [13, 39, 63, 78, 102], i.e.,

$$\mathcal{J}^k \cup \mathcal{J}^{k+1} \cup \cdots \cup \mathcal{J}^{k+T-1} = \mathcal{N}, \quad k = 0, 1, ... \tag{2.6}$$

As we shall see, this generalized Gauss-Seidel rule can also be applied to our nonsmooth separable problem to achieve global convergence. However, for the convergence rate analysis, we need a more restrictive choice of $\mathcal{J}^k$, specifically, there exists a subsequence $\mathcal{T} \subseteq \{0, 1, ...\}$ such that

$$0 \in \mathcal{T}, \qquad \mathcal{N} = \Big(\text{disjoint union of } \mathcal{J}^k, \mathcal{J}^{k+1}, ..., \mathcal{J}^{\tau(k)-1}\Big) \quad \forall k \in \mathcal{T}, \tag{2.7}$$

where $\tau(k) \stackrel{\text{def}}{=} \min\{k' \in \mathcal{T} \mid k' > k\}$. In particular, (2.7) is a special case of (2.6) with $T \leq n$. This choice seems most effective when P is block-separable with large blocks, as in the case of group Lasso (1.4); see [70].

For smooth optimization, $\mathcal{J}^k$ can also be chosen in a *Gauss-Southwell* manner, indexing partial derivatives of the objective function that are within a multiplicative

factor of being maximum in magnitude [35, 78, 91]. This can be extended to our nonsmooth separable problem as follows. We will see in Lemma 2.2 that an $x \in \mathrm{dom} P$ is a stationary point of F_c if and only if $d_H(x;\mathcal{N}) = 0$. Thus, $\|d_H(x;\mathcal{N})\|_\infty$ acts as a scaled "residual" function (with scaling matrix H), measuring how close x comes to being stationary for F_c. Moreover, if H is diagonal, then the separability of P means that $d_H(x;\mathcal{N})_j$, the jth components of $d_H(x;\mathcal{N})$, depends on x_j only and is easily computable.

- If $P \equiv 0$, then $d_H(x;\mathcal{N})_j = -\nabla f(x)_j / H_{jj}$.
- If P is given by (1.2), then $d_H(x;\mathcal{N})_j = \mathrm{mid}\{l_j - x_j, -\nabla f(x)_j/H_{jj}, u_j - x_j\}$.
- If P is the 1-norm, then $d_H(x;\mathcal{N})_j = -\mathrm{mid}\{(\nabla f(x)_j - c)/H_{jj}, x_j, (\nabla f(x)_j + c)/H_{jj}\}$.

[$\mathrm{mid}\{a,b,c\}$ denotes the median (mid-point) of a, b, c.] Accordingly, we choose $\mathcal{J}^k$ to satisfy

$$\|d_{D^k}(x^k;\mathcal{J}^k)\|_\infty \geq \upsilon \|d_{D^k}(x^k;\mathcal{N})\|_\infty, \tag{2.8}$$

where $0 < \upsilon \leq 1$ and $D^k \succ 0_n$ is diagonal (e.g., $D^k = I$ or $D^k = \mathrm{diag}(H^k)$). Other norms beside ∞-norm can also be used. We will call (2.8) the *Gauss-Southwell-r* rule. Notice that $\mathcal{J}^k = \mathcal{N}$ is a valid choice. If P is the indicator function for a closed convex set $X \subseteq \mathbb{R}^n$, then $d_I(x;\mathcal{N}) = [x - \nabla f(x)]_X^+ - x$, where $[x]_X^+$ denotes the orthogonal projection of x onto X. Thus, $d_H(x;\mathcal{N})$ is a generalization of the projection residual used in error bounds and convergence rate analysis of descent methods for constrained smooth optimization [25, 60, 61, 62, 63, 84, 100].

We will see that the above Gauss-Southwell-r rule yields global convergence of the CGD method. However, this rule has thus far resisted a convergence rate analysis. The difficulty lies in that the nonsmooth objective function F_c can have different local growth rates (linear or quadratic) along different coordinate directions, and

this is not adequately captured by the residual $d_{D^k}(x^k;\mathcal{N})$; see Section 2.6 for more discussions. This motivated us to consider a (new) Gauss-Southwell rule based on the optimal objective value of (2.1) rather than the norm of its optimal solution. For any $x \in \mathrm{dom}P$, nonempty $\mathcal{J} \subseteq \mathcal{N}$, and $H \succ 0_n$, define $q_H(x;\mathcal{J})$ to be the optimal objective value of (2.1), i.e.,

$$q_H(x;\mathcal{J}) \stackrel{\mathrm{def}}{=} \left(\nabla f(x)^T d + \frac{1}{2} d^T H d + cP(x+d) - cP(x)\right)_{d=d_H(x;\mathcal{J})}. \tag{2.9}$$

Thus $q_H(x;\mathcal{J})$ estimates the descent in F_c from x to $x+d_H(x;\mathcal{J})$. We have from (2.3) in Lemma 2.1 that $q_H(x;\mathcal{J}) \leq -\frac{1}{2} d_H(x;\mathcal{J})^T H d_H(x;\mathcal{J}) \leq 0$, so that $q_H(x;\mathcal{N}) = 0$ if and only if $d_H(x;\mathcal{N}) = 0$. Thus, like $\|d_H(x;\mathcal{N})\|_\infty$, $-q_H(x;\mathcal{N})$ acts as a "residual" function, measuring how close x comes to being stationary for F_c. If P is separable and H is diagonal, then $q_H(x;\mathcal{J})$ is separable in the sense that $q_H(x;\mathcal{J}) = \sum_{j\in\mathcal{J}} q_H(x;j)$. Accordingly, we choose $\mathcal{J}^k$ to satisfy

$$q_{D^k}(x^k;\mathcal{J}^k) \leq \upsilon\ q_{D^k}(x^k;\mathcal{N}), \tag{2.10}$$

where $0 < \upsilon \leq 1$, $D^k \succ 0_n$ is diagonal (e.g., $D^k = I$ or $D^k = \mathrm{diag}(H^k)$). We call this the *Gauss-Southwell-q* rule. Notice that $\mathcal{J}^k = \mathcal{N}$ is a valid choice. These Gauss-Southwell rules seem most effective when P is separable; see Section 2.6.

2.2 Properties of Search Direction

In this section we study properties of the search direction $d_H(x,\mathcal{J})$ and the residual $d_H(x;\mathcal{N})$ which will be useful for analyzing the global convergence and asymptotic convergence rate of the CGD method.

Formally, we say that $x \in \mathbb{R}^n$ is *a stationary point* of F_c if $x \in \mathrm{dom}F_c$ and $F_c'(x;d) \geq 0$ for all $d \in \mathbb{R}^n$. The following lemma gives an alternative characterization of stationarity that will be often used in our analysis.

Lemma 2.2 *For any $H \succ 0_n$, an $x \in \mathrm{dom}P$ is a stationary point of F_c if and only if $d_H(x;\mathcal{N}) = 0$.*

Proof. Fix any $x \in \mathrm{dom}P$ and $H \succ 0_n$. If $d_H(x;\mathcal{N}) \neq 0$, then (2.2) and (2.3) show that $d_H(x;\mathcal{N})$ is a descent direction for F_c at x, implying that x is not a stationary point of F_c. Conversely, if $d_H(x;\mathcal{N}) = 0$, then

$$g^T u + \frac{1}{2} u^T H u + cP(x+u) - cP(x) \geq 0 \quad \forall u \in \mathbb{R}^n,$$

where $g = \nabla f(x)$. For any $d \in \mathbb{R}^n$, letting $u = \alpha d$ for $\alpha > 0$ yields

$$\alpha g^T d + \frac{1}{2}\alpha^2 d^T H d + cP(x+\alpha d) - cP(x) \geq 0 \quad \forall \alpha > 0. \tag{2.11}$$

Since $f(x+\alpha d) - f(x) = \alpha g^T d + o(\alpha)$, this together with (2.11) yields

$$\begin{aligned} F_c'(x;d) &= \lim_{\alpha \downarrow 0} \frac{f(x+\alpha d) - f(x) + cP(x+\alpha d) - cP(x)}{\alpha} \\ &\geq \lim_{\alpha \downarrow 0} \frac{o(\alpha) - \frac{1}{2}\alpha^2 d^T H d}{\alpha} = 0 \quad \forall d \in \mathbb{R}^n. \end{aligned}$$

Hence $F_c'(x;d) \geq 0$ for all d, implying that x is a stationary point of F_c. ∎

The next lemma shows that $\|d_H(x;\mathcal{J})\|$ changes not too fast with the quadratic coefficients H. It will be used to prove Theorems 2.1, 2.2, 2.3, and 2.4.

Lemma 2.3 *For any $x \in \mathrm{dom}P$, nonempty $\mathcal{J} \subseteq \mathcal{N}$, and $H \succ 0_n$, $\tilde{H} \succ 0_n$, let $d = d_H(x;\mathcal{J})$ and $\tilde{d} = d_{\tilde{H}}(x;\mathcal{J})$. Then*

$$\|\tilde{d}\| \leq \frac{1 + \lambda_{\max}(S) + \sqrt{1 - 2\lambda_{\min}(S) + \lambda_{\max}(S)^2}}{2} \, \frac{\lambda_{\max}(H_{\mathcal{J}\mathcal{J}})}{\lambda_{\min}(\tilde{H}_{\mathcal{J}\mathcal{J}})} \, \|d\|, \tag{2.12}$$

where $S = H_{\mathcal{J}\mathcal{J}}^{-1/2} \tilde{H}_{\mathcal{J}\mathcal{J}} H_{\mathcal{J}\mathcal{J}}^{-1/2}$. If $H_{\mathcal{J}\mathcal{J}} \succ \tilde{H}_{\mathcal{J}\mathcal{J}}$, then also

$$\|d\| \leq \sqrt{\frac{\lambda_{\max}(H_{\mathcal{J}\mathcal{J}} - \tilde{H}_{\mathcal{J}\mathcal{J}})}{\lambda_{\min}(H_{\mathcal{J}\mathcal{J}} - \tilde{H}_{\mathcal{J}\mathcal{J}})}} \, \|\tilde{d}\|. \tag{2.13}$$

Proof. Since $d_j = \tilde{d}_j = 0$ for all $j \notin \mathcal{J}$, it suffices to prove the lemma for the case of $\mathcal{J} = \mathcal{N}$. Let $g = \nabla f(x)$. By the definition of d and $\tilde{d}$ and Fermat's rule [90, Theorem 10.1],

$$\begin{aligned} d &\in \arg\min_u \; (g + Hd)^T u + cP(x+u) - cP(x), \\ \tilde{d} &\in \arg\min_u \; (g + \tilde{H}\tilde{d})^T u + cP(x+u) - cP(x). \end{aligned}$$

Thus

$$\begin{aligned}(g+Hd)^T d + cP(x+d) - cP(x) &\leq (g+Hd)^T \tilde{d} + cP(x+\tilde{d}) - cP(x),\\ (g+\tilde{H}\tilde{d})^T \tilde{d} + cP(x+\tilde{d}) - cP(x) &\leq (g+\tilde{H}\tilde{d})^T d + cP(x+d) - cP(x).\end{aligned}$$

Adding the above two inequalities and rearranging terms yield

$$d^T H d - d^T(H+\tilde{H})\tilde{d} + \tilde{d}^T \tilde{H} \tilde{d} \leq 0.$$

Then, by completing the square on the first two terms, we have

$$\|H^{1/2}d - H^{-1/2}(H+\tilde{H})\tilde{d}/2\|^2 \leq \|H^{-1/2}(H+\tilde{H})\tilde{d}\|^2/4 - \tilde{d}^T\tilde{H}\tilde{d}.$$

By making the substitution $u = H^{1/2}d$, $\tilde{u} = H^{1/2}\tilde{d}$, this can be rewritten as

$$\|u - (I+S)\tilde{u}/2\|^2 \leq \|(I+S)\tilde{u}\|^2/4 - \tilde{u}^T S \tilde{u}.$$

The right-hand side simplifies to $\|(I-S)\tilde{u}\|^2/4$, so taking square root of both sides yields

$$\|u - (I+S)\tilde{u}/2\| \leq \|(I-S)\tilde{u}\|/2.$$

We apply the triangular inequality to the left-hand side and rearrange terms to obtain

$$\|(I+S)\tilde{u}\|/2 - \|(I-S)\tilde{u}\|/2 \leq \|u\|.$$

Multiplying both sides by $2\|(I+S)\tilde{u}\| + 2\|(I-S)\tilde{u}\|$ and simplifying yields

$$4\tilde{u}^T S \tilde{u} \leq 2\|u\|(\|(I+S)\tilde{u}\| + \|(I-S)\tilde{u}\|).$$

Since $S \succ 0_n$, this together with $\|(I+S)\tilde{u}\| \leq (1+\lambda_{\max}(S))\|\tilde{u}\|$ and $\|(I-S)\tilde{u}\| \leq \sqrt{1-2\lambda_{\min}(S)+\lambda_{\max}(S)^2}\|\tilde{u}\|$ yields

$$2\tilde{u}^T S \tilde{u} \leq \|u\|(1+\lambda_{\max}(S) + \sqrt{1-2\lambda_{\min}(S)+\lambda_{\max}(S)^2})\|\tilde{u}\|.$$

Since $\tilde{u}^T S \tilde{u} = \tilde{d}^T \tilde{H} \tilde{d} \geq \lambda_{\min}(\tilde{H})\|\tilde{d}\|^2$ and $\|u\| \leq \sqrt{\lambda_{\max}(H)}\|d\|$, $\|\tilde{u}\| \leq \sqrt{\lambda_{\max}(H)}\|\tilde{d}\|$, this yields (2.12).

Suppose $H \succ \tilde{H}$. From the definition of d and $\tilde{d}$, we have

$$\begin{aligned} g^T d + \frac{1}{2} d^T H d + cP(x+d) - cP(x) &\leq g^T \tilde{d} + \frac{1}{2} \tilde{d}^T H \tilde{d} + cP(x+\tilde{d}) - cP(x), \\ g^T \tilde{d} + \frac{1}{2} \tilde{d}^T \tilde{H} \tilde{d} + cP(x+\tilde{d}) - cP(x) &\leq g^T d + \frac{1}{2} d^T \tilde{H} d + cP(x+d) - cP(x). \end{aligned}$$

Adding the above two inequalities and rearranging terms yields

$$d^T (H - \tilde{H}) d \leq \tilde{d}^T (H - \tilde{H}) \tilde{d}.$$

Hence

$$\lambda_{\min}(H - \tilde{H}) \|d\|^2 \leq \lambda_{\max}(H - \tilde{H}) \|\tilde{d}\|^2,$$

which proves (2.13). ■

If $H = \gamma I$ and $\tilde{H} = \tilde{\gamma} I$ with $\gamma \geq \tilde{\gamma} > 0$, Lemma 2.3 yields that

$$\|d\| \leq \|\tilde{d}\| \leq \frac{\gamma}{\tilde{\gamma}} \|d\|.$$

By switching the roles of H and $\tilde{H}$, (2.12) also yields $\|d\| = O(\|\tilde{d}\|)$. However, this bound seems not as sharp as (2.13). If $\bar{\lambda} I \succeq H \succeq \underline{\lambda} I \succ 0_n$, then $H \succ \frac{\underline{\lambda}}{2} I$, so Lemma 2.3 and the above bound yield

$$\begin{aligned} \|d_H(x; \mathcal{N})\| &\leq \sqrt{\frac{\lambda_{\max}(H - \frac{\underline{\lambda}}{2} I)}{\lambda_{\min}(H - \frac{\underline{\lambda}}{2} I)}} \, \|d_{\frac{\underline{\lambda}}{2} I}(x; \mathcal{N})\| \\ &\leq \sqrt{2\frac{\bar{\lambda}}{\underline{\lambda}} - 1} \|d_{\frac{\underline{\lambda}}{2} I}(x; \mathcal{N})\| \leq \sqrt{2\frac{\bar{\lambda}}{\underline{\lambda}} - 1} \max\{1, \frac{2}{\underline{\lambda}}\} \|d_I(x; \mathcal{N})\| \end{aligned}$$

for all $x \in \mathrm{dom} P$.

The next lemma shows that $d_H(x; \mathcal{J})$ changes not too fast with the linear coefficients $\nabla f(x)$. It will be used to prove Theorem 2.2 on the linear convergence of the CGD method.

Lemma 2.4 *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a smooth function satisfying $(\nabla h(u) - \nabla h(v))^T (u - v) \geq \rho \|u - v\|_p^p$ for all $u, v \in \mathbb{R}^n$, for some $\rho > 0$ and $p > 1$. Let q satisfy $\frac{1}{p} + \frac{1}{q} = 1$.*

Then, for any $x \in \mathrm{dom} P$, *nonempty* $\mathcal{J} \subseteq \mathcal{N}$, *and* $\bar{g}, \tilde{g} \in \mathbb{R}^n$,

$$\|\bar{d} - \tilde{d}\|_p \leq \rho^{-q/p} \|\bar{g}_{\mathcal{J}} - \tilde{g}_{\mathcal{J}}\|_q^{q/p},$$

where $\bar{d} = \underset{d | d_j = 0 \ \forall j \notin \mathcal{J}}{\arg\min} \ \bar{g}^T d + h(d) + cP(x+d) - cP(x)$ *and* $\tilde{d} = \underset{d | d_j = 0 \ \forall j \notin \mathcal{J}}{\arg\min} \ \tilde{g}^T d + h(d) + cP(x+d) - cP(x)$.

Proof. By assumption, h is strictly convex and coercive, so $\bar{d}$ and $\tilde{d}$ are well defined. By Fermat's rule [90, Theorem 10.1],

$$\bar{d} \in \underset{d | d_j = 0 \ \forall j \notin \mathcal{J}}{\arg\min} \ (\bar{g} + \nabla h(\bar{d}))^T d + cP(x+d) - cP(x),$$
$$\tilde{d} \in \underset{d | d_j = 0 \ \forall j \notin \mathcal{J}}{\arg\min} \ (\tilde{g} + \nabla h(\tilde{d}))^T d + cP(x+d) - cP(x).$$

Hence

$$(\bar{g} + \nabla h(\bar{d}))^T \bar{d} + cP(x+\bar{d}) - cP(x) \leq (\bar{g} + \nabla h(\bar{d}))^T \tilde{d} + cP(x+\tilde{d}) - cP(x),$$

$$(\tilde{g} + \nabla h(\tilde{d}))^T \tilde{d} + cP(x+\tilde{d}) - cP(x) \leq (\tilde{g} + \nabla h(\tilde{d}))^T \bar{d} + cP(x+\bar{d}) - cP(x).$$

Summing the above two inequalities and rearranging terms, we have

$$(\tilde{g} - \bar{g})^T (\bar{d} - \tilde{d}) \geq (\nabla h(\bar{d}) - \nabla h(\tilde{d}))^T (\bar{d} - \tilde{d}) \geq \rho \|\bar{d} - \tilde{d}\|_p^p.$$

Since $\bar{d}_j = \tilde{d}_j = 0$ for all $j \notin \mathcal{J}$ and $\|u\|_q \|v\|_p \geq u^T v$ for any $u, v \in \mathbb{R}^n$, this yields

$$\|\tilde{g}_{\mathcal{J}} - \bar{g}_{\mathcal{J}}\|_q \|\bar{d} - \tilde{d}\|_p \geq \rho \|\bar{d} - \tilde{d}\|_p^p,$$

which, upon simplification, proves the desired result. ∎

It can be shown that $h(d) = \frac{1}{p}\|d\|_p^p$, with $p \geq 2$, satisfies the assumption of Lemma 2.4 with $\rho = 1/2^{p-2}$.

We say that P is *block-separable* with respect to nonempty $\mathcal{J} \subseteq \mathcal{N}$ if

$$P(x) = P_{\mathcal{J}}(x_{\mathcal{J}}) + P_{\mathcal{J}_{\mathrm{C}}}(x_{\mathcal{J}_{\mathrm{C}}}) \quad \forall x \in \mathbb{R}^n, \tag{2.14}$$

for some proper, convex, lsc functions $P_{\mathcal{J}}$ and $P_{\mathcal{J}_C}$. In this case, the subproblem (2.1) reduces to the following subproblem:

$$\min_{d_{\mathcal{J}}} \ \nabla f(x)_{\mathcal{J}}^T d_{\mathcal{J}} + \frac{1}{2} d_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} d_{\mathcal{J}} + cP_{\mathcal{J}}(x_{\mathcal{J}} + d_{\mathcal{J}}) \tag{2.15}$$

where $H_{\mathcal{J}\mathcal{J}}$ is the principal submatrix of H indexed by $\mathcal{J}$. Using this observation, we have the next lemma concerning stepsizes satisfying the Armijo descent condition (2.4). This lemma will be used to prove Theorems 2.1(f), 2.2 and 2.4.

Lemma 2.5 *For any $x \in \mathrm{dom} P$, $H \succ 0_n$, and nonempty $\mathcal{J} \subseteq \mathcal{N}$, let $d = d_H(x; \mathcal{J})$ and $g = \nabla f(x)$. For any $\theta \in [0, 1)$, the following results hold with $\Delta = g^T d + \theta d^T H d + cP(x + d) - cP(x)$.*

(a) *If P is block-separable with respect to $\mathcal{J}$, then, for any $\bar{x} \in \mathbb{R}^n$, $\alpha \in (0, 1]$, and $x' = x + \alpha d$,*

$$(g + Hd)_{\mathcal{J}}^T (x' - \bar{x})_{\mathcal{J}} + cP_{\mathcal{J}}(x'_{\mathcal{J}}) - cP_{\mathcal{J}}(\bar{x}_{\mathcal{J}}) \leq (\alpha - 1) \left[(1 - \theta) d^T H d + \Delta \right].$$

(b) *If f satisfies*

$$\|\nabla f(y) - \nabla f(z)\| \leq L \|y - z\| \quad \forall y, z \in \mathrm{dom} P, \tag{2.16}$$

for some $L \geq 0$, and $H \succeq \underline{\lambda} I$, where $\underline{\lambda} > 0$, then the descent condition

$$F_c(x + \alpha d) - F_c(x) \leq \sigma \alpha \Delta, \tag{2.17}$$

is satisfied for any $\sigma \in (0, 1)$ whenever $0 \leq \alpha \leq \min\{1, 2\underline{\lambda}(1 - \sigma + \sigma\theta)/L\}$.

Proof. (a) Since $d = d_H(x; \mathcal{J})$, by (2.15) and Fermat's rule [90, Theorem 10.1],

$$d_{\mathcal{J}} \in \arg\min_{u_{\mathcal{J}}} (g + Hd)_{\mathcal{J}}^T u_{\mathcal{J}} + cP_{\mathcal{J}}(x_{\mathcal{J}} + u_{\mathcal{J}}) - cP_{\mathcal{J}}(x_{\mathcal{J}}).$$

Thus,

$$(g + Hd)_{\mathcal{J}}^T d_{\mathcal{J}} + cP_{\mathcal{J}}(x_{\mathcal{J}} + d_{\mathcal{J}}) - cP_{\mathcal{J}}(x_{\mathcal{J}}) \leq (g + Hd)_{\mathcal{J}}^T (\bar{x} - x)_{\mathcal{J}} + cP_{\mathcal{J}}(\bar{x}_{\mathcal{J}}) - cP_{\mathcal{J}}(x_{\mathcal{J}}). \tag{2.18}$$

Since $x' = x + \alpha d$, we have

$$\begin{aligned}
& (g + Hd)_{\mathcal{J}}^T (x' - \bar{x})_{\mathcal{J}} + cP_{\mathcal{J}}(x'_{\mathcal{J}}) - cP_{\mathcal{J}}(\bar{x}_{\mathcal{J}}) \\
= \; & (\alpha - 1)(g + Hd)_{\mathcal{J}}^T d_{\mathcal{J}} + cP_{\mathcal{J}}(x'_{\mathcal{J}}) + (g + Hd)_{\mathcal{J}}^T (x + d - \bar{x})_{\mathcal{J}} - cP_{\mathcal{J}}(\bar{x}_{\mathcal{J}}) \\
\leq \; & (\alpha - 1)(g + Hd)_{\mathcal{J}}^T d_{\mathcal{J}} + cP_{\mathcal{J}}(x'_{\mathcal{J}}) - cP_{\mathcal{J}}(x_{\mathcal{J}} + d_{\mathcal{J}}) \\
= \; & (\alpha - 1)(g + Hd)^T d + cP(x') - cP(x + d) \\
\leq \; & (\alpha - 1)(g + Hd)^T d + (1 - \alpha)cP(x) + \alpha cP(x + d) - cP(x + d) \\
= \; & (\alpha - 1)(g^T d + d^T H d + cP(x + d) - cP(x)) \\
= \; & (\alpha - 1)(1 - \theta) d^T H d + (\alpha - 1)\Delta,
\end{aligned}$$

where the second step uses (2.18), the third step uses $d_j = 0$ for all $j \notin \mathcal{J}$, and the fourth step uses the convexity of P and $0 < \alpha \leq 1$. This proves the desired result.

(b) For any $\alpha \in [0, 1]$, we have from the convexity of P and the Cauchy-Schwarz inequality that

$$\begin{aligned}
& F_c(x + \alpha d) - F_c(x) \\
= \; & f(x + \alpha d) - f(x) + cP(x + \alpha d) - cP(x) \\
= \; & \alpha \nabla f(x)^T d + cP(x + \alpha d) - cP(x) + \int_0^1 (\nabla f(x + t\alpha d) - \nabla f(x))^T (\alpha d)\, dt \\
\leq \; & \alpha \nabla f(x)^T d + \alpha (cP(x + d) - cP(x)) + \alpha \int_0^1 \|\nabla f(x + t\alpha d) - \nabla f(x)\| \|d\|\, dt \\
\leq \; & \alpha (\nabla f(x)^T d + cP(x + d) - cP(x)) + \alpha^2 \frac{L}{2} \|d\|^2 \\
= \; & \alpha (g^T d + \theta d^T H d + cP(x + d) - cP(x)) - \alpha \gamma d^T H d + \alpha^2 \frac{L}{2} \|d\|^2, \qquad (2.19)
\end{aligned}$$

where the third step uses the convexity of P; the fourth step uses (2.16) and the convexity of $\mathrm{dom} P$, in which x and $x + d$ lie. If $\alpha \leq 2\underline{\lambda}(1 - \sigma + \sigma\theta)/L$, then $d^T H d \geq \underline{\lambda}\|d\|^2$ implies

$$\begin{aligned}
\alpha \frac{L}{2} \|d\|^2 - \theta d^T H d \; & \leq \; (1 - \sigma + \sigma\theta) d^T H d - \theta d^T H d \\
& = \; (1 - \sigma)(1 - \theta) d^T H d \\
& \leq \; -(1 - \sigma)(g^T d + \theta d^T H d + cP(x + d) - cP(x)),
\end{aligned}$$

where the third step uses (2.3) in Lemma 2.1. This together with (2.19) proves (2.17). ■

If P is separable, then P is block-separable with respect to every nonempty $\mathcal{J} \subseteq \mathcal{N}$, with $P_{\mathcal{J}}(x_{\mathcal{J}}) = \sum_{j \in \mathcal{J}} P_j(x_j)$. The converse also holds, since if P is block-separable with respect to $\mathcal{J}, K \subseteq \mathcal{N}$ such that $\mathcal{J} \cap K \neq \emptyset$, then P is block-separable with respect to $\mathcal{J} \cap K$.[1]

2.3 Global Convergence Analysis

In this section we analyze the global convergence of the CGD method under the following reasonable assumption on the choice of H^k. The proof uses Lemmas 2.1, 2.2, 2.3, and 2.5(b).

Assumption 2.1 *$\bar{\lambda} I \succeq H^k \succeq \underline{\lambda} I$ for all k, where $0 < \underline{\lambda} \leq \bar{\lambda}$.*

Theorem 2.1 *Let $\{x^k\}$, $\{d^k\}$, $\{H^k\}$ be sequences generated by the CGD method under Assumption 2.1, where $\{\alpha^k\}$ is chosen by the Armijo rule with $\inf_k \alpha^k_{\text{init}} > 0$. Then the following results hold.*

(a) *$\{F_c(x^k)\}$ is nonincreasing and Δ^k given by (2.5) satisfies*

$$-\Delta^k \geq (1-\theta) d^{kT} H^k d^k \geq (1-\theta)\underline{\lambda} \|d^k\|^2 \quad \forall k, \tag{2.20}$$

$$F_c(x^{k+1}) - F_c(x^k) \leq \sigma \alpha^k \Delta^k \leq 0 \quad \forall k. \tag{2.21}$$

(b) *If $\{x^k\}_{\mathcal{K}}$ is a convergent subsequence of $\{x^k\}$, then $\{\alpha^k \Delta^k\} \to 0$ and $\{d^k\}_{\mathcal{K}} \to 0$. If in addition $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k, where $0 < \underline{\delta} \leq \bar{\delta}$, then $\{d_{D^k}(x^k; \mathcal{J}^k)\}_{\mathcal{K}} \to 0$.*

---

[1] Why? Fix $x_{\mathcal{J}_C} = \bar{x}_{\mathcal{J}_C}$ for some $\bar{x}_{\mathcal{J}_C} \in \text{dom} P_{\mathcal{J}_C}$ and vary $x_{\mathcal{J}}$. Since $P_{\mathcal{J}}(x_{\mathcal{J}}) + P_{\mathcal{J}_C}(\bar{x}_{\mathcal{J}_C}) = P_K(x_K) + P_{K^c}(x_{K^c})$, $P_{\mathcal{J}}(x_{\mathcal{J}})$ is a sum of two functions, one of $x_{\mathcal{J} \cap K}$ only and the other of $x_{\mathcal{J} \setminus K}$ only.

(c) *If* $\{\mathcal{J}^k\}$ *is chosen by the Gauss-Southwell-r rule* (2.8) *and* $\bar{\delta}I \succeq D^k \succeq \underline{\delta}I$ *for all* k, *where* $0 < \underline{\delta} \leq \bar{\delta}$, *then every cluster point of* $\{x^k\}$ *is a stationary point of* F_c.

(d) *If* $\{\mathcal{J}^k\}$ *is chosen by the Gauss-Southwell-q rule* (2.10), $\bar{\delta}I \succeq D^k \succeq \underline{\delta}I$ *for all* k, *where* $0 < \underline{\delta} \leq \bar{\delta}$, *and either (1)* P *is continuous on* $\mathrm{dom}P$ *or (2)* $\inf_k \alpha^k > 0$ *or (3)* $\alpha^k_{\mathrm{init}} = 1$ *for all* k, *then every cluster point of* $\{x^k\}$ *is a stationary point of* F_c.

(e) *If* $\{\mathcal{J}^k\}$ *is chosen by the generalized Gauss-Seidel rule* (2.6), P *is block-separable with respect to* $\mathcal{J}^k$ *for all* k, *and* $\sup_k \alpha^k < \infty$, *then every cluster point of* $\{x^k\}$ *is a stationary point of* F_c.

(f) *If* f *satisfies* (2.16) *for some* $L \geq 0$, *then* $\inf_k \alpha^k > 0$. *If* $\lim_{k\to\infty} F_c(x^k) > -\infty$ *also, then* $\{\Delta^k\} \to 0$ *and* $\{d^k\} \to 0$.

Proof. (a) The inequalities (2.20) follow from (2.5), (2.3) in Lemma 2.1, $0 \leq \theta < 1$, and $H^k \succeq \underline{\lambda}I$. Since $x^{k+1} = x^k + \alpha^k d^k$ and α^k is chosen by the Armijo rule (2.4), we have (2.21) and hence $\{F_c(x^k)\}$ is nonincreasing.

(b) Let $\{x^k\}_{\mathcal{K}}$ be a subsequence of $\{x^k\}$ converging to some $\bar{x}$. Since F_c is lsc, $F_c(\bar{x}) \leq \liminf_{\substack{k\to\infty \\ k\in\mathcal{K}}} F_c(x^k)$. Since $\{F_c(x^k)\}$ is nonincreasing, this implies that $\{F_c(x^k)\}$ converges to a finite limit. Hence, $\{F_c(x^k) - F_c(x^{k+1})\} \to 0$. Then, by (2.21),

$$\{\alpha^k \Delta^k\} \to 0. \tag{2.22}$$

Suppose that $\{d^k\}_{\mathcal{K}} \not\to 0$. By passing to a subsequence if necessary, we can assume that, for some $\delta > 0$, $\|d^k\| \geq \delta$ for all $k \in \mathcal{K}$. Then, by (2.22), $\{\alpha^k\}_{\mathcal{K}} \to 0$. Since $\inf_k \alpha^k_{\mathrm{init}} > 0$, there exists some index $\bar{k} \geq 0$ such that $\alpha^k < \alpha^k_{\mathrm{init}}$ and $\alpha^k \leq \beta$ for all $k \in \mathcal{K}$ with $k \geq \bar{k}$. Since α^k is chosen by the Armijo rule, this implies that

$$F_c(x^k + (\alpha^k/\beta)d^k) - F_c(x^k) > \sigma(\alpha^k/\beta)\Delta^k \quad \forall k \in \mathcal{K},\ k \geq \bar{k}.$$

Thus

$$\begin{aligned}
& \sigma\Delta^k \\
= \; & \sigma\left(g^{k^T}d^k + \theta d^{k^T}H^k d^k + cP(x^k+d^k) - cP(x^k)\right) \\
< \; & \frac{f(x^k+(\alpha^k/\beta)d^k) - f(x^k) + cP(x^k+(\alpha^k/\beta)d^k) - cP(x^k)}{\alpha^k/\beta} \\
\leq \; & \frac{f(x^k+(\alpha^k/\beta)d^k) - f(x^k) + (\alpha^k/\beta)cP(x^k+d^k) + (1-\alpha^k/\beta)cP(x^k) - cP(x^k)}{\alpha^k/\beta} \\
= \; & \frac{f(x^k+(\alpha^k/\beta)d^k) - f(x^k)}{\alpha^k/\beta} + cP(x^k+d^k) - cP(x^k) \quad \forall k \in \mathcal{K},\, k \geq \bar{k},
\end{aligned}$$

where the second inequality uses $0 < \alpha^k/\beta \leq 1$ and the convexity of P. Using the definition of Δ^k, we can rewrite this as

$$-(1-\sigma)\Delta^k + \theta d^{k^T}H^k d^k \leq \frac{f(x^k+(\alpha^k/\beta)d^k) - f(x^k)}{\alpha^k/\beta} - g^{k^T}d^k.$$

Since, by (2.20), the left-hand side is greater than or equal to $((1-\sigma)(1-\theta)+\theta)\underline{\lambda}\|d^k\|^2$, dividing both sides by $\|d^k\|$ yields

$$((1-\sigma)(1-\theta)+\theta)\underline{\lambda}\|d^k\| \leq \frac{f(x^k+\hat{\alpha}^k d^k/\|d^k\|) - f(x^k)}{\hat{\alpha}^k} - \frac{g^{k^T}d^k}{\|d^k\|} \quad \forall k \in \mathcal{K},\, k \geq \bar{k}, \tag{2.23}$$

where we let $\hat{\alpha}^k = \alpha^k\|d^k\|/\beta$. By (2.20), $-\alpha^k\Delta^k \geq (1-\theta)\underline{\lambda}\alpha^k\|d^k\|^2 \geq (1-\theta)\underline{\lambda}\alpha^k\|d^k\|\delta$ for all $k \in \mathcal{K}$, so (2.22) and $(1-\theta)\underline{\lambda} > 0$ imply $\{\alpha^k\|d^k\|\}_{\mathcal{K}} \to 0$ and hence $\{\hat{\alpha}^k\}_{\mathcal{K}} \to 0$. Also, since $\{d^k/\|d^k\|\}_{\mathcal{K}}$ is bounded, by passing to a subsequence if necessary, we can assume that $\{d^k/\|d^k\|\}_{\mathcal{K}} \to$ some $\bar{d}$. Taking the limit as $k \in \mathcal{K}, k \to \infty$ in the inequality (2.23) and using the smoothness of f, we obtain

$$0 < ((1-\sigma)(1-\theta)+\theta)\underline{\lambda}\delta \leq \nabla f(\bar{x})^T\bar{d} - \nabla f(\bar{x})^T\bar{d} \; = \; 0,$$

a clear contradiction. Thus $\{d^k\}_{\mathcal{K}} \to 0$.

Suppose that, in addition, $\bar{\delta}I \succeq D^k \succeq \underline{\delta}I$ for all k. Then, for each k,

$$\frac{\bar{\delta}}{\underline{\lambda}}I \succeq \bar{\delta}(H^k_{\mathcal{J}^k\mathcal{J}^k})^{-1} \succeq (H^k_{\mathcal{J}^k\mathcal{J}^k})^{-1/2}D^k_{\mathcal{J}^k\mathcal{J}^k}(H^k_{\mathcal{J}^k\mathcal{J}^k})^{-1/2} \succeq \underline{\delta}(H^k_{\mathcal{J}^k\mathcal{J}^k})^{-1} \succeq \frac{\underline{\delta}}{\bar{\lambda}}I.$$

Then (2.12) in Lemma 2.3 yields

$$\|d_{D^k}(x^k;\mathcal{J}^k)\| \le \frac{1+\bar{\delta}/\underline{\lambda}+\sqrt{1-2\underline{\delta}/\bar{\lambda}+(\bar{\delta}/\underline{\lambda})^2}}{2}\ \frac{\bar{\lambda}}{\underline{\delta}}\ \|d^k\|. \tag{2.24}$$

Since $\{d^k\}_{\mathcal{K}} \to 0$, this implies $\{d_{D^k}(x^k;\mathcal{J}^k)\}_{\mathcal{K}} \to 0$.

(c) Suppose that $\mathcal{J}^k$ is chosen by the Gauss-Southwell-r rule and $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k. Suppose that $\bar{x}$ is a cluster point of $\{x^k\}$. Let $\{x^k\}_{\mathcal{K}}$ be a subsequence of $\{x^k\}$ converging to $\bar{x}$. Then, by (b), $\{d_{D^k}(x^k;\mathcal{J}^k)\}_{\mathcal{K}} \to 0$. By the Gauss-Southwell-r rule (2.8), this in turn implies $\{r^k\}_{\mathcal{K}} \to 0$, where we denote for simplicity $r^k = d_{D^k}(x^k;\mathcal{N})$. By (2.1), we have

$$\begin{aligned} & g^{k^T} r^k + \frac{1}{2} r^{k^T} D^k r^k + cP(x^k + r^k) - cP(x^k) \\ \le\ & g^{k^T}(x - x^k) + \frac{1}{2}(x - x^k)^T D^k (x - x^k) + cP(x) - cP(x^k) \quad \forall x \in \mathbb{R}^n, \end{aligned}$$

so adding both sides by $cP(x^k)$ and then passing to the limit as $k \in \mathcal{K}, k \to \infty$ and using the smoothness of f and lsc of P yields

$$cP(\bar{x}) \le \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2}(x - \bar{x})^T \bar{D}(x - \bar{x}) + cP(x) \quad \forall x \in \mathbb{R}^n,$$

where $\bar{D}$ is any cluster point of $\{D^k\}_{\mathcal{K}}$. Since $D^k \succeq \underline{\delta} I$ for all $k \in \mathcal{K}$, $\bar{D} \succ 0_n$. This shows that $d_{\bar{D}}(\bar{x};\mathcal{N}) = 0$ so that, by Lemma 2.2, $\bar{x}$ is a stationary point of F_c.

(d) Suppose that $\mathcal{J}^k$ is chosen by the Gauss-Southwell-q rule, and $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k. Suppose that $\bar{x}$ is a cluster point of $\{x^k\}$. Let $\{x^k\}_{\mathcal{K}}$ be a subsequence of $\{x^k\}$ converging to $\bar{x}$. By (b), $\{d^k\}_{\mathcal{K}} \to 0$ and $\{\tilde{d}^k\}_{\mathcal{K}} \to 0$, where we denote $\tilde{d}^k = d_{D^k}(x^k;\mathcal{J}^k)$.

Suppose furthermore that either P is continuous on $\mathrm{dom} P$ or $\alpha^k_{\mathrm{init}} = 1$ for all k or $\inf_k \alpha^k > 0$. We will show that

$$\{q_{D^k}(x^k;\mathcal{J}^k)\}_{\mathcal{K}} \to 0. \tag{2.25}$$

Then, by (2.10), $\{q_{D^k}(x^k;\mathcal{N})\}_{\mathcal{K}} \to 0$. By (2.9), (2.3) in Lemma 2.1, and $D^k \succeq \underline{\delta} I$, we also have

$$q_{D^k}(x^k;\mathcal{N}) \le -\frac{1}{2} d_{D^k}(x^k;\mathcal{N})^T D^k d_{D^k}(x^k;\mathcal{N}) \le -\frac{\underline{\delta}}{2}\|d_{D^k}(x^k;\mathcal{N})\|^2 \quad \forall k, \tag{2.26}$$

this implies that $\{d_{D^k}(x^k;\mathcal{N})\}_{\mathcal{K}} \to 0$. Then, arguing as in the proof of (c), we obtain that $\bar{x}$ is a stationary point of F_c.

We prove (2.25) by contradiction. Suppose that (2.25) is false, i.e.,

$$q_{D^k}(x^k;\mathcal{J}^k) \le -\delta \quad \forall k \in \mathcal{K}', \tag{2.27}$$

for some $\delta > 0$ and $\mathcal{K}' \subseteq \mathcal{K}$ with infinitely many elements. We show below that

$$\{P(x^k + \tilde{d}^k) - P(x^k)\}_{\mathcal{K}'} \to 0. \tag{2.28}$$

Case (1): Suppose P is continuous on $\mathrm{dom}P$. Since $x^k, x^k+\tilde{d}^k \in \mathrm{dom}P$, $\{x^k\}_{\mathcal{K}'} \to \bar{x}$, and $\{\tilde{d}^k\}_{\mathcal{K}'} \to 0$, (2.28) readily follows.

Case (2): Suppose $\inf_k \alpha^k > 0$. By (b), $\{\Delta^k\}_{\mathcal{K}'} \to 0$. We also have from $d^k = d_{H^k}(x^k;\mathcal{J}^k)$ and $\tilde{d}^k = d_{D^k}(x^k;\mathcal{J}^k)$ for all k that

$$\begin{aligned}
\Delta^k + (\frac{1}{2} - \theta){d^k}^T H^k d^k &= {g^k}^T d^k + \frac{1}{2}{d^k}^T H^k d^k + cP(x^k + d^k) - cP(x^k) \\
&\le {g^k}^T \tilde{d}^k + \frac{1}{2}(\tilde{d}^k)^T H^k \tilde{d}^k + cP(x^k + \tilde{d}^k) - cP(x^k) \\
&\le \frac{1}{2}(\tilde{d}^k)^T H^k \tilde{d}^k - (\tilde{d}^k)^T D^k \tilde{d}^k,
\end{aligned}$$

where the last step uses (2.3) in Lemma 2.1. Since $\{d^k\}_{\mathcal{K}'} \to 0$ and $\{H^k\}$ is bounded, the left-hand side tends to zero as $k \in \mathcal{K}', k \to \infty$. Since $\{\tilde{d}^k\}_{\mathcal{K}'} \to 0$ and $\{D^k\}$ is bounded, the right-hand side tends to zero as $k \in \mathcal{K}', k \to \infty$. Thus the quantity between them also tends to zero as $k \in \mathcal{K}', k \to \infty$. Since f is smooth so that $\{g^k\}_{\mathcal{K}'} \to \nabla f(\bar{x})$, (2.28) follows.

Case (3): Suppose $\alpha^k_{\mathrm{init}} = 1$ for all k. By further passing to a subsequence if necessary, we can assume that either $\alpha^k = 1$ for all $k \in \mathcal{K}'$ or $\alpha^k < 1$ for all $k \in \mathcal{K}'$. In the first subcase, the same argument as in Case (2) proves (2.28). In the second subcase, we have from the Armijo rule that $F_c(x^k + d^k) > F_c(x^k) + \sigma\Delta^k$ or, equivalently,

$$f(x^k + d^k) - f(x^k) + (1-\sigma)c(P(x^k + d^k) - P(x^k)) > \sigma({g^k}^T d^k + \theta {d^k}^T H^k d^k)$$

for all $k \in \mathcal{K}'$. Since $\sigma < 1$, $\{x^k\}_{\mathcal{K}'} \to \bar{x}$, $\{d^k\}_{\mathcal{K}'} \to 0$, and $\{H^k\}_{\mathcal{K}'}$ is bounded, this shows that $\liminf_{\substack{k\in\mathcal{K}'\\ k\to\infty}}(P(x^k + d^k) - P(x^k)) \geq 0$. Since

$$0 \geq \Delta^k = g^{k^T} d^k + \theta d^{k^T} H^k d^k + cP(x^k + d^k) - cP(x^k) \quad \forall k,$$

this in turn yields that $\{\Delta^k\}_{\mathcal{K}'} \to 0$. Then, the same argument as in Case (2) proves (2.28) also.

We have from (2.9) that

$$q_{D^k}(x^k; \mathcal{J}^k) = g^{k^T} \tilde{d}^k + \frac{1}{2}(\tilde{d}^k)^T D^k \tilde{d}^k + cP(x^k + \tilde{d}^k) - cP(x^k) \quad \forall k \in \mathcal{K}'.$$

Since f is smooth, $\{x^k\}_{\mathcal{K}'} \to \bar{x}$, $\{\tilde{d}^k\}_{\mathcal{K}'} \to 0$, and $\{D^k\}_{\mathcal{K}'}$ is bounded, this together with (2.28) yields $\{q_{D^k}(x^k; \mathcal{J}^k)\}_{\mathcal{K}'} \to 0$, contradicting (2.27).

(e) Suppose that $\{\mathcal{J}^k\}$ is chosen by the generalized Gauss-Seidel rule (2.6), P is block-separable with respect to $\mathcal{J}^k$ for all k, and $\sup_k \alpha^k < \infty$. The latter implies $\{\alpha^k\}$ is bounded. Suppose that $\bar{x}$ is a cluster point of $\{x^k\}$. Let $\{x^k\}_{\mathcal{K}}$ be a subsequence of $\{x^k\}$ converging to $\bar{x}$. By further passing to a subsequence if necessary, we can assume that $\{H^k\}_{\mathcal{K}} \to$ some $\bar{H}$ and $\mathcal{J}^k = \mathcal{J}$ for all $k \in \mathcal{K}$. Since $H^k \succeq \underline{\lambda} I$ for all k, we have $\bar{H} \succeq \underline{\lambda} I \succ 0_n$. By the definition of d^k and $\mathcal{J}^k = \mathcal{J}$, we have from (2.14) that

$$\begin{aligned} & g^{k^T}_{\mathcal{J}} d^k_{\mathcal{J}} + \frac{1}{2}(d^k_{\mathcal{J}})^T H^k_{\mathcal{J}\mathcal{J}} d^k_{\mathcal{J}} + cP_{\mathcal{J}}(x^k_{\mathcal{J}} + d^k_{\mathcal{J}}) - cP_{\mathcal{J}}(x^k_{\mathcal{J}}) \\ \leq \; & g^{k^T}_{\mathcal{J}} (x - x^k)_{\mathcal{J}} + \frac{1}{2}(x - x^k)^T_{\mathcal{J}} H^k_{\mathcal{J}\mathcal{J}} (x - x^k)_{\mathcal{J}} + cP_{\mathcal{J}}(x_{\mathcal{J}}) - cP_{\mathcal{J}}(x^k_{\mathcal{J}}) \quad \forall x \in \mathbb{R}^n. \end{aligned}$$

Adding both sides by $cP_{\mathcal{J}}(x^k_{\mathcal{J}})$ yields that

$$\begin{aligned} & g^{k^T}_{\mathcal{J}} d^k_{\mathcal{J}} + \frac{1}{2}(d^k_{\mathcal{J}})^T H^k_{\mathcal{J}\mathcal{J}} d^k_{\mathcal{J}} + cP_{\mathcal{J}}(x^k_{\mathcal{J}} + d^k_{\mathcal{J}}) \\ \leq \; & g^{k^T}_{\mathcal{J}} (x - x^k)_{\mathcal{J}} + \frac{1}{2}(x - x^k)^T_{\mathcal{J}} H^k_{\mathcal{J}\mathcal{J}} (x - x^k)_{\mathcal{J}} + cP_{\mathcal{J}}(x_{\mathcal{J}}) \quad \forall x \in \mathbb{R}^n. \end{aligned}$$

Since $\{d^k\}_{\mathcal{K}} \to 0$ by (b), passing to the limit as $k \in \mathcal{K}, k \to \infty$ and using the smoothness of f and lsc of $P_{\mathcal{J}}$ yields

$$cP_{\mathcal{J}}(\bar{x}_{\mathcal{J}}) \leq \nabla f(\bar{x})^T_{\mathcal{J}} (x - \bar{x})_{\mathcal{J}} + \frac{1}{2}(x - \bar{x})^T_{\mathcal{J}} \bar{H}_{\mathcal{J}\mathcal{J}} (x - \bar{x})_{\mathcal{J}} + cP_{\mathcal{J}}(x_{\mathcal{J}}) \quad \forall x \in \mathbb{R}^n.$$

This shows that $d_{\bar{H}}(\bar{x};\mathcal{J}) = 0$ so that, by Lemma 2.2, $\bar{x}$ is a stationary point of F_c with respect to the components indexed by $\mathcal{J}$, i.e., $F_c'(\bar{x};d) \geq 0$ for all $d \in \mathbb{R}^n$ with $d_j = 0$ for $j \notin \mathcal{J}$.

Since $\{d^k\}_{\mathcal{K}} \to 0$, the boundedness of $\{\alpha^k\}_{\mathcal{K}}$ implies $\{x^{k+1}\}_{\mathcal{K}} \to \bar{x}$. This in turn implies $\{d^{k+1}\}_{\mathcal{K}} \to 0$ by (b), and so $\{x^{k+2}\}_{\mathcal{K}} \to \bar{x}$. Continuing in this manner, we obtain that $\{x^{k+\ell}\}_{\mathcal{K}} \to \bar{x}$, for $\ell = 1, ..., T-1$. Thus, we can apply the above argument to $\{x^{k+\ell}\}_{\mathcal{K}}$ to obtain

$$F_c'(\bar{x};d) \geq 0 \quad \forall d \in \mathbb{R}^n \text{ with } d_j = 0 \; \forall j \notin \mathcal{J}_\ell, \quad \ell = 0, 1, ..., T-1,$$

where $\mathcal{J}_0, \mathcal{J}_1, ..., \mathcal{J}_{T-1}$ are nonempty subsets of $\mathcal{N}$ whose union equals $\mathcal{N}$; see (2.6). Since f is differentiable and P is block-separable with respect to $\mathcal{J}_0, \mathcal{J}_1, ..., \mathcal{J}_{T-1}$, this in turn implies that $F_c'(\bar{x};d) \geq 0$ for all $d \in \mathbb{R}^n$, so $\bar{x}$ is a stationary point of F_c.

(f) Since α^k is chosen by the Armijo rule, either $\alpha^k = \alpha^k_{\text{init}}$ or else, by Lemma 2.5(b), $\alpha^k/\beta > \min\{1, 2\underline{\lambda}(1-\sigma+\sigma\theta)/L\}$. Since $\inf_k \alpha^k_{\text{init}} > 0$, this implies $\inf_k \alpha^k > 0$. If $\lim_{k\to\infty} F_c(x^k) > -\infty$ also, then this and (2.21) imply $\{\Delta^k\} \to 0$, which together with (2.20) imply $\{d^k\} \to 0$. ∎

Notice that the assumption $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ in Theorem 2.1(b), (c), (d) is automatically satisfied if we choose $D^k = I$ or $D^k = \text{diag}(H^k)$ under Assumption 2.1. Also, the assumption $\sup_k \alpha^k < \infty$ in Theorem 2.1(e) is automatically satisfied if we choose $\sup_k \alpha^k_{\text{init}} < \infty$. In the case where P is separable, in addition to being proper convex lsc, P is automatically continuous on dom$P$ [90, Corollary 2.37].

To our knowledge, Theorem 2.1 is new even in the unconstrained smooth case (i.e., $P \equiv 0$). Theorem 2.1(e) shows that the CGD method has stronger global convergence properties than the coordinate minimization method when both update coordinates in a Gauss-Seidel manner. In particular, the CGD method cannot cycle on Powell's example [83].

If we choose $\mathcal{J}^k = \mathcal{N}$ and $H^k = \lambda^k I$ for all k with $\bar{\lambda} \geq \lambda^k \geq \underline{\lambda} > 0$, then the CGD method is closely related to the method of Fukushima and Mine [36]. Since $\mathcal{J}^k$

satisfies (2.8), Theorem 2.1(c) implies that every cluster point of $\{x^k\}$ is a stationary point of F_c. In contrast, the convergence result in [36, Theorem 4.1] further assumes that ∇f has a Lipschitz property and $P'(x;\cdot)$ has a continuity property.

2.4 Convergence Rate Analysis

In this section we analyze the asymptotic convergence rate of the CGD method under the following assumption, analogous to that made for constrained smooth optimization [63]. In what follows, $\bar{X}$ denotes the set of stationary points of F_c and

$$\operatorname{dist}(x, \bar{X}) = \min_{\bar{x} \in \bar{X}} \|x - \bar{x}\| \qquad \forall x \in \mathbb{R}^n.$$

Assumption 2.2 **(a)** *$\bar{X} \neq \emptyset$ and, for any $\zeta \geq \min_x F_c(x)$, there exist scalars $\tau > 0$ and $\epsilon > 0$ such that*

$$\operatorname{dist}(x, \bar{X}) \leq \tau \|d_I(x; \mathcal{N})\| \quad \textit{whenever} \quad F_c(x) \leq \zeta,\ \|d_I(x; \mathcal{N})\| \leq \epsilon. \tag{2.29}$$

(b) *There exists a scalar $\delta > 0$ such that*

$$\|x - y\| \geq \delta \quad \textit{whenever} \quad x \in \bar{X},\ y \in \bar{X},\ F_c(x) \neq F_c(y).$$

Assumption 2.2 is a generalization of Assumptions A and B in [63] for constrained smooth problems. Assumption 2.2(a) is a local Lipschitzian error bound assumption, saying that the distance from x to $\bar{X}$ is locally in the order of the norm of the residual at x. Error bounds of this kind have been extensively studied. Assumption 2.2(b) says that the isocost surfaces of F_c restricted to the solution set $\bar{X}$ are "properly separated." Assumption 2.2(b) holds automatically if f is a convex function. It also holds if f is quadratic and P is polyhedral, as can be seen by applying [60, Lemma 3.1] to (1.5).

Our analysis will use ideas from the proof in [63, Appendix] for smooth constrained problems, i.e., P is the indicator function for a nonempty closed convex set. However,

the nonsmooth nature of the objective function F_c requires new proof ideas. In particular, the proof in [63, Appendix] relies on using the error bound (2.29) to derive an inequality like

$$F_c(x^{k+1}) - \bar{v} \leq \tau' \|x^{k+1} - x^k\|^2$$

for all k sufficiently large, where $\tau' > 0$ and $\bar{v} = \lim_{k\to\infty} F_c(x^k)$; see [63, page 175]. For the nonsmooth case, we cannot derive this same inequality but instead work with a weaker inequality whereby the quadratic term $\|x^{k+1} - x^k\|^2$ is replaced by $-\Delta^k$.

We first have the following technical lemma.

Lemma 2.6 *Assume that f satisfies* (2.16) *for some $L \geq 0$. If $\{x^k\}_{\mathcal{K}}$ is a subsequence of a sequence $\{x^k\}$ in $\mathbb{R}^n$ satisfying $\{x^k - \bar{x}^k\}_{\mathcal{K}} \to 0$ and*

$$F_c(\bar{x}^k) = \bar{v} \quad \forall k \in \mathcal{K},\ k \geq \hat{k}, \tag{2.30}$$

for some index $\hat{k}$, $\bar{v} \in \mathbb{R}$, $\mathcal{K} \subseteq \{0, 1, ...\}$, and $\bar{x}^k \in \bar{X}$, then

$$\liminf_{\substack{k\in\mathcal{K}\\ k\to\infty}} F_c(x^k) \geq \bar{v}.$$

Proof. Fix any index $k \in \mathcal{K},\ k \geq \hat{k}$. Since $\bar{x}^k$ is a stationary point of F_c, we have

$$\nabla f(\bar{x}^k)^T(x^k - \bar{x}^k) + cP(x^k) - cP(\bar{x}^k) \geq 0.$$

We also have from the Mean Value Theorem that

$$f(x^k) - f(\bar{x}^k) = \nabla f(\psi^k)^T(x^k - \bar{x}^k),$$

for some ψ^k lying on the line segment joining x^k with $\bar{x}^k$. Since $x^k, \bar{x}^k$ lie in the convex set dom$P$, so does ψ^k. Combining these two relations and using (2.30), we obtain

$$\begin{aligned} \bar{v} - F_c(x^k) &\leq (\nabla f(\bar{x}^k) - \nabla f(\psi^k))^T(x^k - \bar{x}^k) \\ &\leq \|\nabla f(\bar{x}^k) - \nabla f(\psi^k)\|\|x^k - \bar{x}^k\| \\ &\leq L\|x^k - \bar{x}^k\|^2, \end{aligned}$$

where the last inequality uses (2.16), the convexity of dom$P$, and $\|\psi^k - \bar{x}^k\| \le \|x^k - \bar{x}^k\|$. This together with $\{x^k - \bar{x}^k\}_{\mathcal{K}} \to 0$ proves the desired result. ■

The next three theorems establish, under Assumptions 2.1–2.2 and (2.16), the linear rate of convergence of the CGD method using either the restricted Gauss-Seidel rule, the Gauss-Southwell-r or the Gauss-Southwell-q rule to choose $\{\mathcal{J}^k\}$. Their proofs use Theorem 2.1 and Lemmas 2.1, 2.3, 2.4, 2.5(a), 2.6. In what follows, by Q-linear and R-linear convergence, we mean linear convergence in the quotient and the root sense, respectively [78, Chapter 9].

Theorem 2.2 *Assume that f satisfies* (2.16) *for some $L \ge 0$. Let $\{x^k\}$, $\{H^k\}$, $\{d^k\}$ be sequences generated by the CGD method satisfying Assumption 2.1, where $\{\mathcal{J}^k\}$ is chosen by the restricted Gauss-Seidel rule* (2.7) *with $\mathcal{T} \subseteq \{0, 1, ...\}$. Then the following results hold.*

(a) $\|d_I(x^k; \mathcal{N})\| \le \sup_j \alpha^j \, C r^k$ *for all $k \in \mathcal{T}$, where $r^k = \sum_{\ell=k}^{\tau(k)-1} \|d^\ell\|$ and $C > 0$ depends on $n, L, \underline{\lambda}, \bar{\lambda}$.*

(b) *If F_c satisfies Assumption 2.2, P is block-separable with respect to $\mathcal{J}^k$ for all k, and $\{\alpha^k\}$ is chosen by the Armijo rule with $\sup_k \alpha^k_{\rm init} \le 1$ and $\inf_k \alpha^k_{\rm init} > 0$, then either $\{F_c(x^k)\} \downarrow -\infty$ or $\{F_c(x^k)\}_{\mathcal{T}}$ converges at least Q-linearly and $\{x^k\}_{\mathcal{T}}$ converges at least R-linearly.*

Proof. (a) Let $g^k = \nabla f(x^k)$ for all k. For each $k \in \mathcal{T}$, we have from (2.7) that

$$\|d_I(x^k; \mathcal{N})\| = \sqrt{\sum_{\ell=k}^{\tau(k)-1} \|d_I(x^k; \mathcal{J}^\ell)\|^2} \le \sum_{\ell=k}^{\tau(k)-1} \|d_I(x^k; \mathcal{J}^\ell)\|.$$

Since $x^\ell_{\mathcal{J}^\ell} = x^k_{\mathcal{J}^\ell}$, we obtain from Lemma 2.4 with $h(u) = \|u\|^2/2$, $p = 2$, $\rho = 1$, $\mathcal{J} = \mathcal{J}^\ell$, $\bar{d} = d_I(x^k; \mathcal{J}^\ell)$, $\tilde{d} = d_I(x^\ell; \mathcal{J}^\ell)$, $\bar{g} = g^k$, $\tilde{g} = g^\ell$ that

$$\|d_I(x^\ell; \mathcal{J}^\ell) - d_I(x^k; \mathcal{J}^\ell)\| \le \|g^\ell_{\mathcal{J}^\ell} - g^k_{\mathcal{J}^\ell}\| \le L\|x^\ell - x^k\|,$$

where the second inequality uses (2.16) and $x^\ell, x^k \in \mathrm{dom}P$. Combining the above two relations and using triangle inequality yield

$$\begin{aligned}\|d_I(x^k;\mathcal{N})\| &\le \sum_{\ell=k}^{\tau(k)-1}\Big(\|d_I(x^\ell;\mathcal{J}^\ell)\| + L\|x^\ell - x^k\|\Big)\\ &\le \sum_{\ell=k}^{\tau(k)-1}\Big(\tilde{\theta}\|d_{H^\ell}(x^\ell;\mathcal{J}^\ell)\| + L\|x^\ell - x^k\|\Big)\\ &\le \sum_{\ell=k}^{\tau(k)-1}\left(\tilde{\theta}\|d^\ell\| + L\sum_{j=k}^{\ell-1}\alpha^j\|d^j\|\right),\end{aligned}$$

where the second step uses Lemma 2.3 with $H = H^\ell$ and $\tilde{H} = I$, and we denote $\tilde{\theta} = (1 + 1/\underline{\lambda} + \sqrt{1 - 2/\bar{\lambda} + 1/\underline{\lambda}^2})\bar{\lambda}/2$; the last step uses $\|x^\ell - x^k\| = \|\sum_{j=k}^{\ell-1}\alpha^j d^j\| \le \sum_{j=k}^{\ell-1}\alpha^j\|d^j\|$. Since $\tau(k) - k \le n$, this yields the desired result.

(b) By Theorem 2.1(a), $\{F_c(x^k)\}$ is nonincreasing. Thus either $\{F_c(x^k)\} \downarrow -\infty$ or $\lim_{k\to\infty} F_c(x^k) > -\infty$. Suppose the latter. Since α^k is chosen by the Armijo rule with $\inf_k \alpha^k_{\rm init} > 0$, Theorem 2.1(f) implies $\{d^k\} \to 0$. Since $\tau(k) - k \le n$ for all $k \in \mathcal{T}$, this implies that $\{r^k\}_\mathcal{T} \to 0$ and hence, by (a), $\{d_I(x^k;\mathcal{N})\}_\mathcal{T} \to 0$. Since $\{F_c(x^k)\}$ is nonincreasing, this implies that $F_c(x^k) \le F_c(x^0)$ and $\|d_I(x^k;\mathcal{N})\| \le \epsilon$ for all $k \in \mathcal{T}$ with $k \ge$ some $\bar{k}$. Then, by (a) and Assumption 2.2(a), we have

$$\|x^k - \bar{x}^k\| \le \tau' r^k \quad \forall k \in \mathcal{T},\ k \ge \bar{k}, \tag{2.31}$$

where $\tau' > 0$ and $\bar{x}^k \in \bar{X}$ satisfies $\|x^k - \bar{x}^k\| = \mathrm{dist}(x^k, \bar{X})$. Since $\{r^k\}_\mathcal{T} \to 0$, this implies $\{x^k - \bar{x}^k\}_\mathcal{T} \to 0$. Since $\{x^{k+1} - x^k\} = \{\alpha^k d^k\} \to 0$, this and Assumption 2.2(b) imply that $\{\bar{x}^k\}_\mathcal{T}$ eventually settles down at some isocost surface of F_c, i.e., there exist an index $\hat{k} \ge \bar{k}$ and $\bar{v} \in \mathbb{R}$ such that $F_c(\bar{x}^k) = \bar{v}$ for all $k \in \mathcal{T}$ with $k \ge \hat{k}$. Then, by Lemma 2.6 with $\mathcal{K} = \mathcal{T}$,

$$\liminf_{\substack{k\in\mathcal{T}\\ k\to\infty}} F_c(x^k) \ge \bar{v}. \tag{2.32}$$

Fix any $k \in \mathcal{T}$. For $\ell \in \{k, k+1, ..., \tau(k)-1\}$, we have from the Armijo rule (2.4) that

$$F_c(x^{\ell+1}) - F_c(x^\ell) \le \sigma\alpha^\ell\Delta^\ell.$$

Summing this over $\ell = k, k+1, ..., \tau(k)-1$ yields that

$$F_c(x^{\tau(k)}) - F_c(x^k) \leq \sum_{\ell=k}^{\tau(k)-1} \sigma\alpha^\ell \Delta^\ell. \tag{2.33}$$

Also, using (2.14) and letting $\xi^\ell = P_{\mathcal{J}^\ell}(x^{\tau(k)}_{\mathcal{J}^\ell})$, $\bar{\xi}^\ell = P_{\mathcal{J}^\ell}(\bar{x}^k_{\mathcal{J}^\ell})$, we have that, for $k \geq \hat{k}$,

$$\begin{aligned}
& F_c(x^{\tau(k)}) - \bar{v} \\
= \;& f(x^{\tau(k)}) + cP(x^{\tau(k)}) - f(\bar{x}^k) - cP(\bar{x}^k) \\
= \;& \nabla f(\tilde{x}^k)^T(x^{\tau(k)} - \bar{x}^k) + \sum_{\ell=k}^{\tau(k)-1} \left[c\xi^\ell - c\bar{\xi}^\ell\right] \\
= \;& (\nabla f(\tilde{x}^k) - g^k)^T(x^{\tau(k)} - \bar{x}^k) + \sum_{\ell=k}^{\tau(k)-1} \left[(g^k - g^\ell)^T_{\mathcal{J}^\ell}(x^{\ell+1} - \bar{x}^k)_{\mathcal{J}^\ell}\right] \\
& - \sum_{\ell=k}^{\tau(k)-1} (H^\ell d^\ell)^T_{\mathcal{J}^\ell}(x^{\ell+1} - \bar{x}^k)_{\mathcal{J}^\ell} + \sum_{\ell=k}^{\tau(k)-1} \left[\left(g^\ell + H^\ell d^\ell\right)^T_{\mathcal{J}^\ell}(x^{\ell+1} - \bar{x}^k)_{\mathcal{J}^\ell} + c\xi^\ell - c\bar{\xi}^\ell\right] \\
\leq \;& L\|\tilde{x}^k - x^k\|\|x^{\tau(k)} - \bar{x}^k\| + \sum_{\ell=k}^{\tau(k)-1} L\|x^k - x^\ell\|\|x^{\ell+1} - \bar{x}^k\| \\
& + \sum_{\ell=k}^{\tau(k)-1} \bar{\lambda}\|d^\ell\|\|x^{\ell+1} - \bar{x}^k\| + \sum_{\ell=k}^{\tau(k)-1} \left[\left(g^\ell + H^\ell d^\ell\right)^T_{\mathcal{J}^\ell}(x^{\ell+1} - \bar{x}^k)_{\mathcal{J}^\ell} + c\xi^\ell - c\bar{\xi}^\ell\right] \\
\leq \;& L\|\tilde{x}^k - x^k\|\|x^{\tau(k)} - \bar{x}^k\| + \sum_{\ell=k}^{\tau(k)-1} L\|x^\ell - x^k\|\|x^{\ell+1} - \bar{x}^k\| \\
& + \sum_{\ell=k}^{\tau(k)-1} \bar{\lambda}\|d^\ell\|\|x^{\ell+1} - \bar{x}^k\| + \sum_{\ell=k}^{\tau(k)-1} (\alpha^\ell - 1)\left[(1-\theta)d^{\ell T} H^\ell d^\ell + \Delta^\ell\right],
\end{aligned} \tag{2.34}$$

where the second step uses the Mean Value Theorem with $\tilde{x}^k$ a point lying on the segment joining $x^{\tau(k)}$ with $\bar{x}^k$; the third step uses (2.7) and $x^{\tau(k)}_{\mathcal{J}^\ell} = x^{\ell+1}_{\mathcal{J}^\ell}$ for $k \leq \ell < \tau(k)$; the fourth step uses $\bar{\lambda} I \succeq H^\ell \succ 0_n$, (2.16), and the convexity of dom$P$; and the last step uses $\xi^\ell = P_{\mathcal{J}^\ell}(x^{\ell+1}_{\mathcal{J}^\ell})$, $\alpha^k \leq \alpha^k_{\text{init}} \leq 1$, and Lemma 2.5(a).

Fix any $k \in \mathcal{T},\ k \geq \hat{k}$. Using the inequalities $\|\tilde{x}^k - x^k\| \leq \|x^{\tau(k)} - x^k\| + \|x^k - \bar{x}^k\|$, $\|x^{\ell+1} - \bar{x}^k\| \leq \|x^{\ell+1} - x^k\| + \|x^k - \bar{x}^k\|$, $\|x^{\ell+1} - x^k\| \leq \sum_{j=k}^{\ell} \alpha^j \|d^j\|$ for $k \leq \ell < \tau(k)$, we see from (2.31) and $\alpha^j \leq 1$ that the right-hand side of (2.34) is bounded above by

$$C_1 \sum_{\ell=k}^{\tau(k)-1} \|d^\ell\|^2 + \sum_{\ell=k}^{\tau(k)-1} (\alpha^\ell - 1)\left[(1-\theta)d^{\ell T} H^\ell d^\ell + \Delta^\ell\right]$$

for some constant $C_1 > 0$ depending on $L, \tau', n, \bar{\lambda}$ only. Since, by (2.20), we have $-\Delta^\ell \geq (1-\theta) d^{\ell T} H^\ell d^\ell \geq (1-\theta)\underline{\lambda}\|d^\ell\|^2$, the above quantity is bounded above by

$$-C_2 \sum_{\ell=k}^{\tau(k)-1} \Delta^\ell$$

for some constant $C_2 > 0$ depending on $L, \tau', n, \bar{\lambda}, \underline{\lambda}, \theta$ only. Combining this with (2.33), (2.34), and $\inf_k \alpha^k > 0$ (see Theorem 2.1(f)) yields

$$F_c(x^{\tau(k)}) - \bar{v} \leq C_3(F_c(x^k) - F_c(x^{\tau(k)})) \quad \forall k \in \mathcal{T},\ k \geq \hat{k},$$

where $C_3 = C_2/(\sigma \inf_k \alpha^k)$. Upon rearranging terms and using (2.32), we have

$$0 \leq F_c(x^{\tau(k)}) - \bar{v} \leq \frac{C_3}{1+C_3}(F_c(x^k) - \bar{v}) \quad \forall k \in \mathcal{T},\ k \geq \hat{k},$$

so $\{F_c(x^k)\}_{\mathcal{T}}$ converges to $\bar{v}$ at least Q-linearly.

Finally, (2.20) implies $\Delta^\ell \leq (\theta - 1)\underline{\lambda}\|d^\ell\|^2$, so that (2.33) and $x^{\ell+1} - x^\ell = \alpha^\ell d^\ell$ yield

$$\sigma(1-\theta)\underline{\lambda} \sum_{\ell=k}^{\tau(k)-1} \frac{\|x^{\ell+1} - x^\ell\|^2}{\alpha^\ell} \leq F_c(x^k) - F_c(x^{\tau(k)}) \quad \forall k \in \mathcal{T},\ k \geq \hat{k}.$$

This implies

$$\begin{aligned} \|x^{\tau(k)} - x^k\| &\leq \sqrt{(\tau(k) - k) \sum_{\ell=k}^{\tau(k)-1} \|x^{\ell+1} - x^\ell\|^2} \\ &\leq \sqrt{n \frac{(\sup_\ell \alpha^\ell)}{\sigma(1-\theta)\underline{\lambda}}(F_c(x^k) - F_c(x^{\tau(k)}))} \quad \forall k \in \mathcal{T},\ k \geq \hat{k}. \end{aligned}$$

Since $\{F_c(x^k) - F_c(x^{\tau(k)})\}_{\mathcal{T}} \to 0$ at least R-linearly and $\sup_\ell \alpha^\ell \leq 1$, this implies that $\{x^k\}_{\mathcal{T}}$ converges at least R-linearly.[2] ∎

---

[2]More precisely, writing $\mathcal{T} = \{k_1, k_2, \ldots\}$, we have $\|x^{k_{t+1}} - x^{k_t}\| = O\left(\sqrt{F_c(x^{k_t}) - F_c(x^{k_{t+1}})}\right) = O(\vartheta^t)$ for $t = 1, 2, \ldots$, where $\vartheta = \sqrt{\frac{C_3}{1+C_3}}$. Thus $\{x^{k_t}\}_{t=1,2,\ldots}$ satisfies Cauchy's criterion for convergence, implying it has a unique limit $\bar{x}$. Moreover, for any $t' > t$, we have $\|x^{k_t} - x^{k_{t'}}\| \leq \sum_{j=t}^{t'-1} \|x^{k_{j+1}} - x^{k_j}\| \leq O(\sum_{j=t}^{t'-1} \vartheta^j) = O(\vartheta^t)$. Taking $t' \to \infty$ yields $\|x^{k_t} - \bar{x}\| = O(\vartheta^t)$ for any t, so $\limsup_{t\to\infty} \|x^{k_t} - \bar{x}\|^{1/t} \leq \vartheta < 1$.

Theorem 2.3 *Suppose that f satisfy* (2.16) *for some $L \geq 0$ and f and P are separable. Let $\{x^k\}$, $\{H^k\}$, $\{d^k\}$ be sequences generated by the CGD method satisfying Assumption 2.1, where $\{\mathcal{J}^k\} = \{j^k\}$ is chosen by Gauss-Southwell-r rule* (2.8) *with $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k ($0 < \underline{\delta} \leq \bar{\delta}$). If F_c satisfies Assumption 2.2 and $\{\alpha^k\}$ is chosen by the Armijo rule with $\sup_k \alpha^k_{\rm init} \leq 1$ and $\inf_k \alpha^k_{\rm init} > 0$, then either $\{F_c(x^k)\} \downarrow -\infty$ or $\{x^k\}$ converges at least R-linearly.*

Proof. By Theorem 2.1(a), $\{F_c(x^k)\}$ is nonincreasing. Thus either $\{F_c(x^k)\} \downarrow -\infty$ or $\lim_{k\to\infty} F_c(x^k) > -\infty$. Suppose the latter. Since α^k is chosen by the Armijo rule with $\inf_k \alpha^k_{\rm init} > 0$, Theorem 2.1(f) implies $\inf_k \alpha^k > 0$, $\{\Delta^k\} \to 0$, and $\{d^k\} \to 0$. This together with (2.8) and (2.24) yields $\{d_{D^k}(x^k; \mathcal{N})\} \to 0$.

By Lemma 2.3 with $\mathcal{J} = \mathcal{N}$, $H = D^k$ and $\tilde{H} = I$, we have

$$\|d_I(x^k; \mathcal{N})\| \leq \frac{1 + 1/\underline{\delta} + \sqrt{1 - 2/\bar{\delta} + (1/\underline{\delta})^2}}{2} \bar{\delta} \, \|d_{D^k}(x^k; \mathcal{N})\| \quad \forall k. \tag{2.35}$$

Hence $\{d_I(x^k; \mathcal{N})\} \to 0$. Since $\{F_c(x^k)\}$ is nonincreasing, this implies that $F_c(x^k) \leq F_c(x^0)$ and $\|d_I(x^k; \mathcal{N})\| \leq \epsilon$ for all $k \geq$ some $\bar{k}$. Then, by Assumption 2.2(a), there exist $\bar{k}$ and $\tau > 0$ such that

$$\|x^k - \bar{x}^k\| \leq \tau \|d_I(x^k; \mathcal{N})\| \quad \forall k \geq \bar{k}, \tag{2.36}$$

where $\bar{x}^k \in \bar{X}$ satisfies $\|x^k - \bar{x}^k\| = \mathrm{dist}(x^k, \bar{X})$. Since $\{d_I(x^k; \mathcal{N})\} \to 0$, this implies $\{x^k - \bar{x}^k\} \to 0$. Since $\{x^{k+1} - x^k\} = \{\alpha^k d^k\} \to 0$, this and Assumption 2.2(b) and the separability of F_c imply that for each $\mathcal{J}$, $\{\bar{x}^k_{\mathcal{J}}\}$ eventually settles down at some isocost surface of $F_{c\mathcal{J}}$, i.e., there exist an index $\hat{k} \geq \bar{k}$ and a scalar $\bar{v}(\mathcal{J})$ such that $F_{c\mathcal{J}}(\bar{x}^k_{\mathcal{J}}) = \bar{v}(\mathcal{J}) \quad \forall k \geq \hat{k}$. By Lemma 2.6 with $\mathcal{K} = \{0, 1, ...\}$ and $F_c = F_{c\mathcal{J}}$,

$$\liminf_{k\to\infty} F_{c\mathcal{J}}(x^k_{\mathcal{J}}) \geq \bar{v}(\mathcal{J}). \tag{2.37}$$

Fix any k, we have from the Armijo rule (2.4) and the separability of F_c that

$$F_c(x^{k+1}) - F_c(x^k) = F_{c\mathcal{J}^k}(x^{k+1}_{\mathcal{J}^k}) - F_{c\mathcal{J}^k}(x^k_{\mathcal{J}^k}) \leq \sigma \alpha^k \Delta^k. \tag{2.38}$$

Also, letting $\mathcal{J} = \mathcal{J}^k$, we have

$$
\begin{aligned}
& F_{c\mathcal{J}}(x_{\mathcal{J}}^{k+1}) - \bar{v}(\mathcal{J}) \qquad (2.39)\\
= \;& f_{\mathcal{J}}(x_{\mathcal{J}}^{k+1}) + cP_{\mathcal{J}}(x_{\mathcal{J}}^{k+1}) - f_{\mathcal{J}}(\bar{x}_{\mathcal{J}}^{k}) - cP_{\mathcal{J}}(\bar{x}_{\mathcal{J}}^{k})\\
= \;& (\nabla f_{\mathcal{J}}(\tilde{x}_{\mathcal{J}}^{k}) - \nabla f_{\mathcal{J}}(x_{\mathcal{J}}^{k}))^T(x_{\mathcal{J}}^{k+1} - \bar{x}_{\mathcal{J}}^{k})\\
& +\nabla f_{\mathcal{J}}(x_{\mathcal{J}}^{k})^T(x_{\mathcal{J}}^{k+1} - \bar{x}_{\mathcal{J}}^{k}) + cP_{\mathcal{J}}(x_{\mathcal{J}}^{k+1}) - cP_{\mathcal{J}}(\bar{x}_{\mathcal{J}}^{k})\\
= \;& (\nabla f_{\mathcal{J}}(\tilde{x}_{\mathcal{J}}^{k}) - \nabla f_{\mathcal{J}}(x_{\mathcal{J}}^{k}))^T(x_{\mathcal{J}}^{k+1} - \bar{x}_{\mathcal{J}}^{k}) - (H^k d^k)_{\mathcal{J}}^T(x^{k+1} - \bar{x}^k)_{\mathcal{J}}\\
& +(\nabla f_{\mathcal{J}}(x_{\mathcal{J}}^{k}) + (H^k d^k)_{\mathcal{J}})^T(x^{k+1} - \bar{x}^k)_{\mathcal{J}} + cP_{\mathcal{J}}(x_{\mathcal{J}}^{k+1}) - cP_{\mathcal{J}}(\bar{x}_{\mathcal{J}}^{k})\\
\leq \;& L\|\tilde{x}_{\mathcal{J}}^{k} - x_{\mathcal{J}}^{k}\|\|x_{\mathcal{J}}^{k+1} - \bar{x}_{\mathcal{J}}^{k}\| + \|(H^k d^k)_{\mathcal{J}}\|\|(x^{k+1} - \bar{x}^k)_{\mathcal{J}}\|\\
& +(\nabla f_{\mathcal{J}}(x_{\mathcal{J}}^{k}) + (H^k d^k)_{\mathcal{J}})^T(x^{k+1} - \bar{x}^k)_{\mathcal{J}} + cP_{\mathcal{J}}(x_{\mathcal{J}}^{k+1}) - cP_{\mathcal{J}}(\bar{x}_{\mathcal{J}}^{k})\\
\leq \;& L\|\tilde{x}_{\mathcal{J}}^{k} - x_{\mathcal{J}}^{k}\|\|x_{\mathcal{J}}^{k+1} - \bar{x}_{\mathcal{J}}^{k}\| + \bar{\lambda}\|d^k\|\|(x^{k+1} - \bar{x}^k)_{\mathcal{J}}\|\\
& +(\alpha^k - 1)(1-\theta){d^k}^T H^k d^k + (\alpha^k - 1)\Delta^k\\
\leq \;& L\|\tilde{x}_{\mathcal{J}}^{k} - x_{\mathcal{J}}^{k}\|\|x_{\mathcal{J}}^{k+1} - \bar{x}_{\mathcal{J}}^{k}\| + \bar{\lambda}\|d^k\|\|(x^{k+1} - \bar{x}^k)_{\mathcal{J}}\|\\
& +\alpha^k\bar{\lambda}\|d^k\|^2 + (\alpha^k - 1)\Delta^k, \qquad (2.40)
\end{aligned}
$$

where the second step comes from using the Mean Value Theorem with some vector $\tilde{x}_{\mathcal{J}^k}$ lying on the segment joining $x_{\mathcal{J}^{k+1}}$ with $\bar{x}_{\mathcal{J}^k}$; the fifth step uses $\bar{\lambda}I \succeq H^k \succ 0_n$ and Lemma 2.5(a); the last step uses $\theta < 1$ and $\bar{\lambda}I \succeq H^k \succ 0_n$.

Using the inequalities $\|\tilde{x}_{\mathcal{J}}^{k} - x_{\mathcal{J}}^{k}\| \leq \|x_{\mathcal{J}}^{k+1} - x_{\mathcal{J}}^{k}\| + \|x_{\mathcal{J}}^{k} - \bar{x}_{\mathcal{J}}^{k}\|$, $\|x_{\mathcal{J}}^{k+1} - \bar{x}_{\mathcal{J}}^{k}\| \leq \|x_{\mathcal{J}}^{k+1} - x_{\mathcal{J}}^{k}\| + \|x_{\mathcal{J}}^{k} - \bar{x}_{\mathcal{J}}^{k}\|$, $\|x_{\mathcal{J}}^{k+1} - x_{\mathcal{J}}^{k}\| = \alpha^k\|d^k\|$, we see from (2.36) and $\sup_k \alpha^k \leq 1$ (since $\sup_k \alpha_{\text{init}}^k \leq 1$) that the right-hand side of (2.40) is bounded above by

$$C_1(\|d^k\| + \|d_I(x^k; \mathcal{N})\|)^2 + (\alpha^k - 1)\Delta^k$$

for some constant $C_1 > 0$ depending on $L, \tau, \bar{\lambda}$ only. By (2.8), (2.24), and (2.35), the above quantity is bounded above by

$$C_2\|d^k\|^2 + (\alpha^k - 1)\Delta^k$$

for some constant $C_2 > 0$ depending on $L, \tau, \bar{\lambda}, \underline{\lambda}, \bar{\delta}, \underline{\delta}, \upsilon$ only. Since, by (2.20), we have $-\Delta^\ell \geq (1-\theta) d^{\ell T} H^\ell d^\ell \geq (1-\theta)\underline{\lambda}\|d^\ell\|^2$, the above quantity is bounded above by

$$-C_3 \Delta^k \tag{2.41}$$

for some constant $C_3 > 0$ depending on $L, \tau, \bar{\lambda}, \underline{\lambda}, \bar{\delta}, \underline{\delta}, \upsilon, \theta$ only. Combining this with (2.38), (2.40), and $\inf_k \alpha^k > 0$ (see Theorem 2.1(f)) yields

$$F_{c\mathcal{J}}(x^{k+1}_{\mathcal{J}}) - \bar{v}(\mathcal{J}) \leq C_4 (F_{c\mathcal{J}}(x^k_{\mathcal{J}}) - F_{c\mathcal{J}}(x^{k+1}_{\mathcal{J}})) \quad \forall k \geq \hat{k},$$

for some $C_4 > 0$. Upon rearranging terms and using (2.37), we have

$$0 \leq F_{c\mathcal{J}}(x^{k+1}_{\mathcal{J}}) - \bar{v}(\mathcal{J}) \leq \frac{C_4}{1+C_4} (F_{c\mathcal{J}}(x^k_{\mathcal{J}}) - \bar{v}(\mathcal{J})) \quad \forall k \geq \hat{k},$$

so

$$F_{c\mathcal{J}}(x^k_{\mathcal{J}}) - \bar{v}(\mathcal{J}) \leq \left(\frac{C_4}{1+C_4}\right)^{\sharp k - 1} (F_{c\mathcal{J}}(x^{\hat{k}}_{\mathcal{J}}) - \bar{v}(\mathcal{J})),$$

where $\sharp k = |\{h \mid \hat{k} \leq h \leq k, \ \mathcal{J} \text{ is chosen at } h\}|$.
Also, by (2.20) and (2.38),

$$\sigma(1-\theta)\underline{\lambda}\|d^k\|^2 \leq F_{c\mathcal{J}}(x^k_{\mathcal{J}}) - F_{c\mathcal{J}}(x^{k+1}_{\mathcal{J}}) \quad \forall k \geq \hat{k}.$$

This together with (2.37) implies $\sigma(1-\theta)\underline{\lambda}\|d^k\|^2 \leq F_{c\mathcal{J}}(x^k_{\mathcal{J}}) - \bar{v}(\mathcal{J}) \quad \forall k \geq \hat{k}$. Hence

$$\|d^k\| \leq \sqrt{\frac{1}{\sigma(1-\theta)\underline{\lambda}} \left(\frac{C_4}{1+C_4}\right)^{\sharp k - 1} (F_{c\mathcal{J}}(x^{\hat{k}}_{\mathcal{J}}) - \bar{v}(\mathcal{J}))}.$$

Let $k(\mathcal{J})$ be the iteration in which the subset $\mathcal{J}$ is chosen. Since the choice for $\mathcal{J}$ is finite, there exists a $\bar{\mathcal{J}}$ such that

$$\limsup_{k \to \infty} \sqrt[k]{\|d^{k(\bar{\mathcal{J}})}\|} < 1.$$

By (2.24),

$$\limsup_{k \to \infty} \sqrt[k]{\|d_{D^{k(\bar{\mathcal{J}})}}(x^{k(\bar{\mathcal{J}})}; \bar{\mathcal{J}})\|} < 1. \tag{2.42}$$

Let $h(\bar{\mathcal{J}})(\geq \hat{k})$ be the iteration in which the subset $\bar{\mathcal{J}}$ is chosen and let $h(\mathcal{J})(\geq \hat{k})$ be the smallest iteration in which the subset $\mathcal{J}(\neq \bar{\mathcal{J}})$ is updated and $h(\mathcal{J}) > h(\bar{\mathcal{J}})$. Then, by (2.8) and $\|\cdot\|_\infty \leq \|\cdot\| \leq \sqrt{n}\|\cdot\|_\infty$,

$$\|d_{D^{h(\mathcal{J})}}(x^{h(\mathcal{J})};\mathcal{J})\| = \|d_{D^{h(\bar{\mathcal{J}})}}(x^{h(\bar{\mathcal{J}})};\mathcal{J})\| \leq \frac{\sqrt{n}\|d_{D^{h(\bar{\mathcal{J}})}}(x^{h(\bar{\mathcal{J}})};\bar{\mathcal{J}})\|}{\upsilon}.$$

This together with (2.42) yields

$$\limsup_{k\to\infty} \sqrt[k]{\|d_{D^{h(\mathcal{J})}}(x^{h(\mathcal{J})};\mathcal{J})\|} < 1. \tag{2.43}$$

Since $\bar{\delta}I \succeq D^k \succeq \underline{\delta}I$ and $\bar{\lambda}I \succeq H^k \succeq \underline{\lambda}I$, we have from (2.12) in Lemma 2.3 that

$$\begin{aligned}
\|d^{k(\mathcal{J})}\| &\leq \frac{1+1/\bar{\lambda}+\sqrt{1-2/\underline{\lambda}+1/\bar{\lambda}^2}}{2}\,\frac{1}{\underline{\lambda}}\,\|d_I(x^{k(\mathcal{J})};\mathcal{J})\|,\\
\|d_I(x^{k(\mathcal{J})};\mathcal{J})\| &\leq \frac{1+1/\underline{\delta}+\sqrt{1-2/\bar{\delta}+1/\underline{\delta}^2}}{2}\,\bar{\delta}\,\|d_{D^{k(\mathcal{J})}}(x^{k(\mathcal{J})};\mathcal{J})\|.
\end{aligned}$$

By the above two inequalities and (2.43), $\limsup_{k\to\infty} \sqrt[k]{\|d^{k(\mathcal{J})}\|} < 1$. Since the choice for $\mathcal{J}$ is finite, $\limsup_{k\to\infty} \sqrt[k]{\|d^k\|} < 1$. This together with $\|x^{k+1}-x^k\| = \alpha^k\|d^k\|$ and $\sup_k \alpha^k \leq 1$ implies that $\{x^k\}$ converges at least R-linearly. ■

Theorem 2.4 *Assume that f satisfies* (2.16) *for some $L \geq 0$. Let $\{x^k\}$, $\{H^k\}$, $\{d^k\}$ be sequences generated by the CGD method satisfying Assumption 2.1, where $\{\mathcal{J}^k\}$ is chosen by Gauss-Southwell-q rule* (2.10) *with P block-separable with respect to $\mathcal{J}^k$ and $\bar{\delta}I \succeq D^k \succeq \underline{\delta}I$ for all k ($0 < \underline{\delta} \leq \bar{\delta}$). If F_c satisfies Assumption 2.2 and $\{\alpha^k\}$ is chosen by the Armijo rule with $\sup_k \alpha^k_{\rm init} \leq 1$ and $\inf_k \alpha^k_{\rm init} > 0$, then either $\{F_c(x^k)\} \downarrow -\infty$ or $\{F_c(x^k)\}$ converges at least Q-linearly and $\{x^k\}$ converges at least R-linearly.*

Proof. For each $k = 0, 1, \ldots$, (2.5) and $d^k = d_{H^k}(x^k;\mathcal{J}^k)$ imply that

$$\begin{aligned}
\Delta^k + \left(\frac{1}{2}-\theta\right) d^{k\,T} H^k d^k &= \nabla f(x^k)^T d^k + \frac{1}{2} d^{k\,T} H^k d^k + cP(x^k+d^k) - cP(x^k)\\
&\leq \nabla f(x^k)^T \tilde{d}^k + \frac{1}{2}(\tilde{d}^k)^T H^k \tilde{d}^k + cP(x^k+\tilde{d}^k) - cP(x^k)\\
&= q_{D^k}(x^k;\mathcal{J}^k) + \frac{1}{2}(\tilde{d}^k)^T (H^k - D^k)\tilde{d}^k\\
&\leq q_{D^k}(x^k;\mathcal{J}^k) + \omega\|d^k\|^2,
\end{aligned} \tag{2.44}$$

where we let $\tilde{d}^k = d_{D^k}(x^k; \mathcal{J}^k)$, and the last step uses (2.24) and $(\tilde{d}^k)^T(H^k - D^k)\tilde{d}^k \leq (\bar{\lambda} - \underline{\delta})\|\tilde{d}^k\|^2$. Here, ω is a constant depending on $\bar{\lambda}, \underline{\lambda}, \bar{\delta}, \underline{\delta}$ only.

By Theorem 2.1(a), $\{F_c(x^k)\}$ is nonincreasing. Thus either $\{F_c(x^k)\} \downarrow -\infty$ or $\lim_{k\to\infty} F_c(x^k) > -\infty$. Suppose the latter. Since α^k is chosen by the Armijo rule with $\inf_k \alpha^k_{\rm init} > 0$, Theorem 2.1(f) implies $\inf_k \alpha^k > 0$, $\{\Delta^k\} \to 0$, and $\{d^k\} \to 0$. Since $\{H^k\}$ is bounded by Assumption 2.1, we obtain from (2.44) that $0 \leq \lim_{k\to\infty} \inf q_{D^k}(x^k; \mathcal{J}^k)$. This together with (2.10) and (2.26) yields $\{d_{D^k}(x^k; \mathcal{N})\} \to 0$.

By Lemma 2.3 with $H = D^k$ and $\tilde{H} = I$,

$$\|d_I(x^k; \mathcal{N})\| \leq \frac{1 + 1/\underline{\delta} + \sqrt{1 - 2/\bar{\delta} + (1/\underline{\delta})^2}}{2}\ \bar{\delta}\ \|d_{D^k}(x^k; \mathcal{N})\| \quad \forall k.$$

Hence $\{d_I(x^k; \mathcal{N})\} \to 0$. Since $\{F_c(x^k)\}$ is nonincreasing, this implies that $F_c(x^k) \leq F_c(x^0)$ and $\|d_I(x^k; \mathcal{N})\| \leq \epsilon$ for all $k \geq$ some $\bar{k}$. Then, by Assumption 2.2(a), we have

$$\|x^k - \bar{x}^k\| \leq \tau \|d_I(x^k; \mathcal{N})\| \quad \forall k \geq \bar{k}, \tag{2.45}$$

where $\tau > 0$ and $\bar{x}^k \in \bar{X}$ satisfies $\|x^k - \bar{x}^k\| = \text{dist}(x^k, \bar{X})$. Since $\{d_I(x^k; \mathcal{N})\} \to 0$, this implies $\{x^k - \bar{x}^k\} \to 0$. Since $\{x^{k+1} - x^k\} = \{\alpha^k d^k\} \to 0$, this and Assumption 2.2(b) imply that $\{\bar{x}^k\}$ eventually settles down at some isocost surface of F_c, i.e., there exist an index $\hat{k} \geq \bar{k}$ and a scalar $\bar{v}$ such that $F_c(\bar{x}^k) = \bar{v}$ for all $k \geq \hat{k}$. By Lemma 2.6 with $\mathcal{K} = \{0, 1, ...\}$,

$$\liminf_{k\to\infty} F_c(x^k) \geq \bar{v}. \tag{2.46}$$

Fix any $k \geq \hat{k}$. Letting $\mathcal{J} = \mathcal{J}^k$ and $\hat{d}^k = d_{D^k}(x^k; \mathcal{J}^k_{\rm C})$, we have from (2.14) that

$$\begin{aligned}
& F_c(x^{k+1}) - \bar{v} \\
= \ & f(x^{k+1}) + cP(x^{k+1}) - f(\bar{x}^k) - cP(\bar{x}^k) \\
= \ & \nabla f(\tilde{x}^k)^T(x^{k+1} - \bar{x}^k) + cP_{\mathcal{J}}(x^{k+1}_{\mathcal{J}}) + cP_{\mathcal{J}_{\rm C}}(x^k_{\mathcal{J}_{\rm C}}) - cP_{\mathcal{J}}(\bar{x}^k_{\mathcal{J}}) - cP_{\mathcal{J}_{\rm C}}(\bar{x}^k_{\mathcal{J}_{\rm C}}) \\
= \ & (\nabla f(\tilde{x}^k) - \nabla f(x^k))^T(x^{k+1} - \bar{x}^k) \\
& -(H^k d^k)^T_{\mathcal{J}}(x^{k+1} - \bar{x}^k)_{\mathcal{J}} - (D^k \hat{d}^k)^T_{\mathcal{J}_{\rm C}}(x^k - \bar{x}^k)_{\mathcal{J}_{\rm C}}
\end{aligned}$$

$$
\begin{aligned}
& +(\nabla f(x^k) + H^k d^k)_{\mathcal{J}}^T (x^{k+1} - \bar{x}^k)_{\mathcal{J}} + cP_{\mathcal{J}}(x_{\mathcal{J}}^{k+1}) - cP_{\mathcal{J}}(\bar{x}_{\mathcal{J}}^k) \\
& +(\nabla f(x^k) + D^k \hat{d}^k)_{\mathcal{J}_{\mathrm{C}}}^T (x^k - \bar{x}^k)_{\mathcal{J}_{\mathrm{C}}} - cP_{\mathcal{J}_{\mathrm{C}}}(\bar{x}_{\mathcal{J}_{\mathrm{C}}}^k) + cP_{\mathcal{J}_{\mathrm{C}}}(x_{\mathcal{J}_{\mathrm{C}}}^k) \\
\leq\ & L\|\tilde{x}^k - x^k\|\|x^{k+1} - \bar{x}^k\| + \|H^k d^k\|\|x^{k+1} - \bar{x}^k\| + \|D^k \hat{d}^k\|\|x^k - \bar{x}^k\| \\
& +(\alpha^k - 1)(1-\theta) {d^k}^T H^k d^k + (\alpha^k - 1)\Delta^k \\
& +(\nabla f(x^k) + D^k \hat{d}^k)_{\mathcal{J}_{\mathrm{C}}}^T (x^k - \bar{x}^k)_{\mathcal{J}_{\mathrm{C}}} - cP_{\mathcal{J}_{\mathrm{C}}}(\bar{x}_{\mathcal{J}_{\mathrm{C}}}^k) + cP_{\mathcal{J}_{\mathrm{C}}}(x_{\mathcal{J}_{\mathrm{C}}}^k) \\
\leq\ & L\|\tilde{x}^k - x^k\|\|x^{k+1} - \bar{x}^k\| + \bar{\lambda}\|d^k\|\|x^{k+1} - \bar{x}^k\| + \bar{\delta}\|\hat{d}^k\|\|x^k - \bar{x}^k\| \\
& +\alpha^k \bar{\lambda}\|d^k\|^2 + (\alpha^k - 1)\Delta^k \\
& -(\nabla f(x^k) + D^k \hat{d}^k)_{\mathcal{J}_{\mathrm{C}}}^T \hat{d}_{\mathcal{J}_{\mathrm{C}}}^k - cP_{\mathcal{J}_{\mathrm{C}}}(x_{\mathcal{J}_{\mathrm{C}}}^k + \hat{d}_{\mathcal{J}_{\mathrm{C}}}^k) + cP_{\mathcal{J}_{\mathrm{C}}}(x_{\mathcal{J}_{\mathrm{C}}}^k) \\
=\ & L\|\tilde{x}^k - x^k\|\|x^{k+1} - \bar{x}^k\| + \bar{\lambda}\|d^k\|\|x^{k+1} - \bar{x}^k\| + \bar{\delta}\|\hat{d}^k\|\|x^k - \bar{x}^k\| \\
& +\alpha^k \bar{\lambda}\|d^k\|^2 + (\alpha^k - 1)\Delta^k - q_{D^k}(x^k; \mathcal{J}_{\mathrm{C}}) - \frac{1}{2}(\hat{d}^k)^T D^k \hat{d}^k, \qquad (2.47)
\end{aligned}
$$

where the second step uses the Mean Value Theorem with $\tilde{x}^k$ a point lying on the segment joining x^{k+1} with $\bar{x}^k$; the third step uses $x_{\mathcal{J}_{\mathrm{C}}}^{k+1} = x_{\mathcal{J}_{\mathrm{C}}}^k$; the fourth step uses (2.16), the convexity of $\mathrm{dom} P$, $\alpha^k \leq \alpha_{\mathrm{init}}^k \leq 1$, and Lemma 2.5(a); the fifth step uses Lemma 2.5(a) (applied to x^k, D^k, $\mathcal{J}_{\mathrm{C}}$, and $\alpha = 1$) as well as $\bar{\lambda} I \succeq H^k \succ 0_n$, $\bar{\delta} I \succeq D^k \succ 0_n$, $\theta < 1$; the last step uses $\hat{d}^k = d_{D^k}(x^k; \mathcal{J}_{\mathrm{C}})$, (2.9), and (2.14).

Using the inequalities $\|\tilde{x}^k - x^k\| \leq \|x^{k+1} - x^k\| + \|x^k - \bar{x}^k\|$, $\|x^{k+1} - \bar{x}^k\| \leq \|x^{k+1} - x^k\| + \|x^k - \bar{x}^k\|$ and $\|x^{k+1} - x^k\| = \alpha^k \|d^k\|$, we see from (2.45), $D^k \succ 0_n$, and $\sup_k \alpha^k \leq 1$ that the right-hand side of (2.47) is bounded above by

$$
C_1(\|d^k\| + \|\hat{d}^k\| + \|d_I(x^k; \mathcal{N})\|)^2 + (\alpha^k - 1)\Delta^k - q_{D^k}(x^k; \mathcal{J}_{\mathrm{C}}^k) \qquad (2.48)
$$

for all $k \geq \hat{k}$, where $C_1 > 0$ is some constant depending on $L, \tau, \bar{\lambda}, \bar{\delta}$ only. Since $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ and $\bar{\lambda} I \succeq H^k \succeq \underline{\lambda} I$, we have from (2.12) in Lemma 2.3 that

$$
\begin{aligned}
\|d_I(x^k; \mathcal{J}^k)\| &\leq \frac{1 + 1/\underline{\lambda} + \sqrt{1 - 2/\bar{\lambda} + 1/\underline{\lambda}^2}}{2}\, \bar{\lambda}\, \|d^k\|, \\
\|d_I(x^k; \mathcal{J}_{\mathrm{C}}^k)\| &\leq \frac{1 + 1/\underline{\delta} + \sqrt{1 - 2/\bar{\delta} + 1/\underline{\delta}^2}}{2}\, \bar{\delta}\, \|\hat{d}^k\|.
\end{aligned}
$$

Thus the quantity in (2.48) is bounded above by

$$C_2\|d^k\|^2 + C_2\|\hat{d}^k\|^2 + (\alpha^k - 1)\Delta^k - q_{D^k}(x^k; \mathcal{J}_{\rm C}^k) \tag{2.49}$$

for all $k \geq \hat{k}$, where $C_2 > 0$ is some constant depending on $L, \tau, \bar{\lambda}, \underline{\lambda}, \bar{\delta}, \underline{\delta}$ only.

By (2.20), we have

$$\underline{\lambda}\|d^k\|^2 \leq {d^k}^T H^k d^k \leq -\frac{1}{1-\theta}\Delta^k \quad \forall k. \tag{2.50}$$

Similarly, by (2.3) in Lemma 2.1 and (2.9), we have $q_{D^k}(x^k; \mathcal{J}_{\rm C}^k) \leq -\frac{1}{2}(\hat{d}^k)^T D^k \hat{d}^k \leq 0$, so that

$$\underline{\delta}\|\hat{d}^k\|^2 \leq (\hat{d}^k)^T D^k \hat{d}^k \leq -2\ q_{D^k}(x^k; \mathcal{J}_{\rm C}^k).$$

Thus, the quantity in (2.49) is bounded above by

$$C_3\ (-\Delta^k - q_{D^k}(x^k; \mathcal{J}_{\rm C}^k)) \tag{2.51}$$

for all $k \geq \hat{k}$, where $C_3 > 0$ is some constant depending on $L, \tau, \bar{\lambda}, \underline{\lambda}, \bar{\delta}, \underline{\delta}, \theta$ only.

By using (2.10) and the block-separability of P and block-diagonal structure of D^k with respect to $\mathcal{J}^k$, we have

$$q_{D^k}(x^k; \mathcal{J}^k) \leq \upsilon\ q_{D^k}(x^k; \mathcal{N}) = \upsilon \left(q_{D^k}(x^k; \mathcal{J}^k) + q_{D^k}(x^k; \mathcal{J}_{\rm C}^k) \right),$$

implying

$$\upsilon\ q_{D^k}(x^k; \mathcal{J}_{\rm C}^k) \geq (1-\upsilon)\ q_{D^k}(x^k; \mathcal{J}^k). \tag{2.52}$$

Combining (2.44) with (2.50) yields

$$\begin{aligned} -q_{D^k}(x^k; \mathcal{J}^k) &\leq -\Delta^k + \left(\theta - \frac{1}{2}\right) {d^k}^T H^k d^k + \omega\|d^k\|^2 \\ &\leq -\Delta^k - \max\left\{0, \theta - \frac{1}{2}\right\} \frac{1}{1-\theta}\Delta^k - \frac{\omega}{\underline{\lambda}(1-\theta)}\Delta^k. \end{aligned} \tag{2.53}$$

Combining (2.52) and (2.53), we see that the quantity in (2.51) is bounded above by

$$-C_4\Delta^k$$

for all $k \geq \hat{k}$, where $C_4 > 0$ is some constant depending on $L, \tau, \bar{\lambda}, \underline{\lambda}, \bar{\delta}, \underline{\delta}, \theta, \upsilon$ only. Thus the right-hand side of (2.47) is bounded above by $-C_4\Delta^k$ for all $k \geq \hat{k}$. Combining this with (2.21), (2.47), and $\inf_k \alpha^k > 0$ (see Theorem 2.1(f)) yields

$$F_c(x^{k+1}) - \bar{v} \leq C_5(F_c(x^k) - F_c(x^{k+1})) \quad \forall k \geq \hat{k},$$

where $C_5 = C_4/(\sigma \inf_k \alpha^k)$. Upon rearranging terms and using (2.46), we have

$$0 \leq F_c(x^{k+1}) - \bar{v} \leq \frac{C_5}{1 + C_5}(F_c(x^k) - \bar{v}) \quad \forall k \geq \hat{k},$$

so $\{F_c(x^k)\}$ converges to $\bar{v}$ at least Q-linearly.

Finally, by (2.21), (2.50), and $x^{k+1} - x^k = \alpha^k d^k$, we have

$$\sigma(1-\theta)\underline{\lambda}\frac{\|x^{k+1} - x^k\|^2}{\alpha^k} \leq F_c(x^k) - F_c(x^{k+1}) \quad \forall k \geq \hat{k}.$$

This implies

$$\|x^{k+1} - x^k\| \leq \sqrt{\frac{\alpha^k}{\sigma(1-\theta)\underline{\lambda}}(F_c(x^k) - F_c(x^{k+1}))} \quad \forall k \geq \hat{k}.$$

Since $\{F_c(x^k) - F_c(x^{k+1})\} \to 0$ at least R-linearly and $\sup_k \alpha^k \leq 1$, this implies that $\{x^k\}$ converges at least R-linearly. ■

The assumption (2.16) in Theorems 2.2, 2.3, and 2.4 can be relaxed to ∇f being Lipschitz continuous on $\mathrm{dom}P \cap (X^0 + \varrho B)$ for some $\varrho > 0$, where B denotes the unit Euclidean ball in $\mathbb{R}^n$ and X^0 denotes the convex hull of the level set $\{x \mid F_c(x) \leq F_c(x^0)\}$. For simplicity, we did not consider this more relaxed assumption.

As we noted in Section 2.1, we have been unable to establish the local linear convergence of the CGD method using the Gauss-Southwell-r rule to choose $\{\mathcal{J}^k\}$. Only in the simple case where f and P are separable have we been able to prove local linear convergence. In fact, even in this case our proof is nontrivial, even though the problem decomposes into n univariate problems. This is because different coordinates can converge at different rates, which is explicitly taken into account in the proof.

2.5 Error Bound

In this section we show that Assumption 2.2(a) is satisfied under problem assumptions analogous to those for constrained smooth optimization. In fact, we will show that error bound for (1.3) is closely related to that for constrained smooth optimization problems.

By using $\mathrm{epi}P = \{(x,\xi)|P(x) \le \xi\}$, we can reformulate (1.1) as the constrained smooth optimization problem (see (1.5)):

$$\min_{(x,\xi)} \ \{ \ f(x) + c\xi \ \mid \ (x,\xi) \in \mathrm{epi}P \ \}. \tag{2.54}$$

For any $(x,\xi) \in \mathrm{epi}P$, the corresponding projection residual is the optimal solution of the subproblem:

$$\min_{(d,\delta)} \ \left\{ \ \nabla f(x)^T d + \frac{1}{2}\|d\|^2 + \frac{1}{2}\delta^2 + c\delta \ \mid \ (x+d, \xi+\delta) \in \mathrm{epi}P \ \right\}. \tag{2.55}$$

The following lemma shows that if P is Lipschitz continuous on $\mathrm{dom}P$, then the norm of this projection residual is bounded above by a multiple of $\|d_I(x;\mathcal{N})\|$ whenever $\xi = P(x)$.

Lemma 2.7 *Suppose that P is Lipschitz continuous on* $\mathrm{dom}P$. *There exists a scalar $\kappa > 0$ (depending only on the Lipschitz constant of P) such that, for any $x \in \mathrm{dom}P$ and $\xi = P(x)$,*

$$\|(\tilde{d}, \tilde{\delta})\| \le \kappa \|d_I(x;\mathcal{N})\|,$$

where $(\tilde{d}, \tilde{\delta})$ is an optimal solution of the subproblem (2.55).

Proof. Fix any $x \in \mathrm{dom}P$ and $\xi = P(x)$. By (2.1), $(d_I(x;\mathcal{N}), \bar{\delta})$ is the optimal solution of the subproblem:

$$\min_{(d,\delta)} \ \left\{ \ \nabla f(x)^T d + \frac{1}{2}\|d\|^2 + c\delta \ \mid \ (x+d, \xi+\delta) \in \mathrm{epi}P \ \right\},$$

where we let $\bar{\delta} = P(x + d_I(x;\mathcal{N})) - P(x)$. By Fermat's rule [90, Theorem 10.1],

$$(d_I(x;\mathcal{N}), \bar{\delta}) \ \in \ \underset{(d,\delta)}{\arg\min} \left\{ \ (\nabla f(x) + d_I(x;\mathcal{N}))^T d + c\delta \ \mid \ (x+d, \xi+\delta) \in \mathrm{epi} P \ \right\}.$$

Hence

$$(\nabla f(x) + d_I(x;\mathcal{N}))^T d_I(x;\mathcal{N}) + c\bar{\delta} \leq (\nabla f(x) + d_I(x;\mathcal{N}))^T \tilde{d} + c\tilde{\delta}.$$

Also, since $(\tilde{d}, \tilde{\delta})$ is the optimal solution of the subproblem (2.55), we have

$$\nabla f(x)^T \tilde{d} + \frac{1}{2}\|\tilde{d}\|^2 + \frac{1}{2}\tilde{\delta}^2 + c\tilde{\delta} \leq \nabla f(x)^T d_I(x;\mathcal{N}) + \frac{1}{2}\|d_I(x;\mathcal{N})\|^2 + \frac{1}{2}\bar{\delta}^2 + c\bar{\delta}.$$

Adding the above two inequalities and simplifying yield

$$\frac{1}{2}\|d_I(x;\mathcal{N})\|^2 - d_I(x;\mathcal{N})^T \tilde{d} + \frac{1}{2}\|\tilde{d}\|^2 + \frac{1}{2}\tilde{\delta}^2 \leq \frac{1}{2}\bar{\delta}^2.$$

Multiplying both sides by 2 and rewriting the first three terms into a square, we have

$$\|d_I(x;\mathcal{N}) - \tilde{d}\|^2 + \tilde{\delta}^2 \leq \bar{\delta}^2.$$

Thus $\tilde{\delta}^2 \leq \bar{\delta}^2$ and $\|d_I(x;\mathcal{N}) - \tilde{d}\|^2 \leq \bar{\delta}^2$. Taking square root of both sides and using the triangle inequality yield

$$|\tilde{\delta}| \leq |\bar{\delta}|, \qquad \|\tilde{d}\| - \|d_I(x;\mathcal{N})\| \leq |\bar{\delta}|. \tag{2.56}$$

Now, the Lipschitz continuity of P on $\mathrm{dom} P$ implies that $|\bar{\delta}| = |P(x + d_I(x;\mathcal{N})) - P(x)| \leq K\|d_I(x;\mathcal{N})\|$, where K is the Lipschitz constant. Then (2.56) yields that

$$|\tilde{\delta}| \leq K\|d_I(x;\mathcal{N})\|, \qquad \|\tilde{d}\| \leq (K+1)\|d_I(x;\mathcal{N})\|,$$

which proves the desired result. ■

The following local error bound results from [60, 61, 62, 84] show that, for all x sufficiently close to $\bar{X}$, $\mathrm{dist}(x, \bar{X})$ can be bounded from above by the norm of the solution of the subproblem (2.55) under certain problem assumptions.

Lemma 2.8 *Assume that $\bar{X} \neq \emptyset$ and any of the following conditions hold.*

C1 *f is quadratic. P is polyhedral.*

C2 *$f(x) = g(Ex) + q^T x$ for all $x \in \mathbb{R}^n$, where $E \in \mathbb{R}^{m \times n}$, $q \in \mathbb{R}^n$, and g is a strongly convex differentiable function on $\mathbb{R}^m$ with ∇g Lipschitz continuous on $\mathbb{R}^m$. P is polyhedral.*

C3 *$f(x) = \max_{y \in Y}\{(Ex)^T y - g(y)\} + q^T x$ for all $x \in \mathbb{R}^n$, where Y is a polyhedral set in $\mathbb{R}^m$, $E \in \mathbb{R}^{m \times n}$, $q \in \mathbb{R}^n$, and g is a strongly convex differentiable function on $\mathbb{R}^m$ with ∇g Lipschitz continuous on $\mathbb{R}^m$. P is polyhedral.*

Then, for any $\zeta \in \mathbb{R}$, there exist scalars $\tau' > 0$ and $\epsilon' > 0$ such that

$$\mathrm{dist}(x, \bar{X}) \leq \tau' \|(\tilde{d}, \tilde{\delta})\| \quad \textit{whenever} \quad F_c(x) \leq \zeta,\ \|(\tilde{d}, \tilde{\delta})\| \leq \epsilon', \tag{2.57}$$

where $(\tilde{d}, \tilde{\delta})$ is the optimal solution of the subproblem (2.55) with $\xi = P(x)$.

Proof. Since epi$P$ is convex, each stationary point $(\bar{x}, \bar{\xi})$ of (2.54) satisfies

$$\nabla f(\bar{x})^T (x - \bar{x}) + c(\xi - \bar{\xi}) \geq 0 \quad \forall (x, \xi) \in \mathrm{epi}P,$$

from which it readily follows that $\bar{\xi} = P(\bar{x})$ and $\bar{x} \in \bar{X}$. Under C1, the objective function of (2.54) is quadratic and epi$P$ is a polyhedral set. Fix any $\zeta \in \mathbb{R}$. By applying [60, Theorem 2.3] (also see [84]) to (2.54), there exist scalars $\tau' > 0$ and $\epsilon' > 0$ such that

$$\min_{\bar{x} \in \bar{X}} \|(x, P(x)) - (\bar{x}, P(\bar{x}))\| \ \leq\ \tau' \|(\tilde{d}, \tilde{\delta})\| \quad \text{whenever} \quad F_c(x) \leq \zeta,\ \|(\tilde{d}, \tilde{\delta})\| \leq \epsilon',$$

where $(\tilde{d}, \tilde{\delta})$ is the optimal solution of (2.55) with $\xi = P(x)$. Since $\|x - \bar{x}\| \leq \|(x, P(x)) - (\bar{x}, P(\bar{x}))\|$ for all $\bar{x} \in \bar{X}$, this proves (2.57).

Under C2, the objective function of (2.54) has the form $g\left([E\ 0] \begin{bmatrix} x \\ \xi \end{bmatrix}\right) + [q^T\ c] \begin{bmatrix} x \\ \xi \end{bmatrix}$ and epi$P$ is a polyhedral set. Then, by applying [61, Theorem 2.1] to (2.54) and

arguing similarly as above, (2.57) can be proved. Under C3, a similar argument using [62, Theorem 4.1] (also see [63, Theorem 2.1]) proves (2.57). ■

By using Lemmas 2.7 and 2.8, we obtain the main result of this section.

Theorem 2.5 *Assumption 2.2(a) is satisfied if $\bar{X} \neq \emptyset$ and any of the conditions C1, C2, C3 in Lemma 2.8 holds or if the following condition holds.*

C4 *f is strongly convex and satisfies* (2.16) *for some $L \geq 0$.*

Proof. Under C1 or C2 or C3, P is polyhedral so, by Example 9.35 in [90], P is Lipschitz continuous on dom$P$. Then Lemmas 2.7 and 2.8 yield that Assumption 2.2(a) holds.

Under C4, for any $x \in \text{dom}P$, since $d_I(x;\mathcal{N})$ is a solution of the subproblem (2.1) with $H = I$, by Fermat's rule [90, Theorem 10.1],

$$d_I(x;\mathcal{N}) \in \arg\min_d \ (\nabla f(x) + d_I(x;\mathcal{N}))^T d + cP(x+d) - cP(x).$$

Hence, for any $\bar{x} \in \bar{X}$ (in fact, $\bar{X}$ is a singleton), we have

$$\begin{aligned} & (\nabla f(x) + d_I(x;\mathcal{N}))^T d_I(x;\mathcal{N}) + cP(x + d_I(x;\mathcal{N})) - cP(x) \\ \leq \ & (\nabla f(x) + d_I(x;\mathcal{N}))^T (\bar{x} - x) + cP(\bar{x}) - cP(x). \end{aligned}$$

Since $\bar{x}$ is a stationary point of F_c, we also have

$$cP(\bar{x}) \leq \nabla f(\bar{x})^T (x + d_I(x;\mathcal{N}) - \bar{x}) + cP(x + d_I(x;\mathcal{N})).$$

Adding the above two inequalities and simplifying yield

$$(\nabla f(x) - \nabla f(\bar{x}))^T (x - \bar{x}) + \|d_I(x;\mathcal{N})\|^2 \leq (\nabla f(\bar{x}) - \nabla f(x))^T d_I(x;\mathcal{N}) + d_I(x;\mathcal{N})^T (\bar{x} - x).$$

It follows from the strong convexity of f and (2.16) that

$$\lambda \|x - \bar{x}\|^2 + \|d_I(x;\mathcal{N})\|^2 \leq L\|x - \bar{x}\| \|d_I(x;\mathcal{N})\| + \|x - \bar{x}\| \|d_I(x;\mathcal{N})\|,$$

for some scalar constants $0 < \lambda \leq L$. Thus

$$\lambda \|x - \bar{x}\|^2 \leq (L+1)\|x - \bar{x}\| \|d_I(x; \mathcal{N})\|.$$

Dividing both sides by $\lambda\|x - \bar{x}\|$ whenever $x \neq \bar{x}$ shows that Assumption 2.2(a) is satisfied with $\tau = (L+1)/\lambda$ and $\epsilon = \infty$ (independent of ζ). ■

Notice that the objective function of (2.54) is not strongly convex under C4. Thus existing error bound results for strongly convex objective function (e.g., [25, Proposition 6.3.1]) cannot be applied to (2.54).

2.6 Implementation and Numerical Experience

In order to better understand its practical performance, we have implemented the CGD method in Matlab, using Matlab's vector operations, to solve the ℓ_1-regularized problem (1.3). In this section, we describe the implementation, together with convergence acceleration techniques, and report our numerical experience on test problems with $n = 1000$ from Moré et al. [73] and the CUTEr set [38]. In particular, we compare the performance of the CGD method using either the Gauss-Seidel rule or the Gauss-Southwell-r rule or the Gauss-Southwell-q rule, with or without acceleration. We also reformulate the ℓ_1-regularized test problems as bound-constrained smooth optimization problems and solve them using the well-known Fortran codes MINOS [75] for constrained smooth optimization and L-BFGS-B [109] for large-scale bound-constrained smooth optimization.

2.6.1 Test functions

For the function f in (1.3), we chose 10 test functions with $n = 1000$ from the set of nonlinear least square functions used by Moré et al. [73]. These functions, listed in Table 2.1, were chosen for their diverse characteristics: convex or nonconvex, sparse or dense Hessian, well-conditioned or ill-conditioned Hessian. Two functions

Table 2.1: Nonlinear least square test functions from [73, pages 26–28].

Name	n	**Description**
BAL	1000	Brown almost-linear function, nonconvex, with dense Hessian.
BT	1000	Broyden tridiagonal function, nonconvex, with sparse Hessian.
DBV	1000	Discrete boundary value function, nonconvex, with sparse Hessian.
ER	1000	Extended Rosenbrook function, nonconvex, with sparse Hessian.
TRIG	1000	Trigonometric function, nonconvex, with dense Hessian.
EPS	1000	Extended Powell singular function, convex, with sparse Hessian.
LR1	1000	$f(x) = \sum_{i=1}^{n} \left(i \left(\sum_{j=1}^{n} j x_j \right) - 1 \right)^2$, convex, with dense Hessian.
LR1Z	1000	$f(x) = \sum_{i=2}^{n-1} \left((i-1) \left(\sum_{j=2}^{n-1} j x_j \right) - 1 \right)^2 + 2$, convex, with dense Hessian.
LFR	1000	$f(x) = \sum_{i=1}^{n} \left(x_i - \frac{2}{n+1} \sum_{j=1}^{n} x_j - 1 \right)^2 + \left(\frac{2}{n+1} \sum_{j=1}^{n} x_j + 1 \right)^2$, strongly convex, with dense Hessian.
VD	1000	Variably dimensioned function $f(x) = \sum_{i=1}^{n} (x_i - 1)^2 + \left(\sum_{i=1}^{n} i(x_i - 1) \right)^2 + \left(\sum_{i=1}^{n} i(x_i - 1) \right)^4$, strongly convex, with dense Hessian.

ER and EPS have block-diagonal Hessians. Since we wish to see how solution sparsity (i.e., number of nonzeros) changes with c, we modified the Extended Powell singular function slightly, replacing "$5^{1/2}(x_{4i-1} - x_{4i})$" with "$5^{1/2}(x_{4i-1} - x_{4i} - 1)$" so that the solution is not always at the origin. We coded the function, gradient, and Hessian diagonals in Matlab using vector operations.

We also chose 10 functions with $n = 1000$ from the unconstrained problems in the CUTEr set [38]. These functions, listed in Table 2.2, were similarly chosen for their diverse characteristics, as well as Hessian availability. The function, gradient, and (sparse) Hessian are called within Matlab using the CUTEr tools "ufr", "ugr" and "ush".

Table 2.2: CUTEr test functions [38].

Name	n	**Description**
EG2	1000	A nonconvex function, with sparse Hessian.
EXTROSNB	1000	The extended Rosenbrock function (nonseparable version), nonconvex, with sparse Hessian.
INDEF	1000	A nonconvex function which is a combination of quadratic and trigonometric functions, with sparse Hessian.
LIARWHD	1000	A simplified version of the NONDIA (Shanno's nondiagonal extension of Rosenbrock function), nonconvex, with sparse Hessian.
NONCVXU2	1000	A nonconvex function with a unique minimum value, with sparse Hessian.
PENALTY1	1000	$f(x) = \sum_{i=1}^{n} 10^{-5}(x_i - 1)^2 + \left(\left(\sum_{j=1}^{n} x_j^2\right) - \frac{1}{4}\right)^2$, nonconvex, with dense Hessian.
WOODS	1000	The extended Woods function, nonconvex, with sparse Hessian.
QUARTC	1000	A simple quartic function, convex, with sparse Hessian.
DIXON3DQ	1000	Dixon's quadratic function, strongly convex, with tridiagonal Hessian.
TRIDIA	1000	Shanno's TRIDIA quadratic function, strongly convex, with tridiagonal Hessian.

2.6.2 Implementation of the CGD method

In our implementation of the CGD method, we choose a diagonal Hessian approximation

$$H^k = \text{diag}\left[\min\{\max\{\nabla^2 f(x^k)_{jj}, 10^{-2}\}, 10^9\}\right]_{j=1,\ldots,n},$$

which has the advantage that d^k has a closed form and can be computed efficiently in Matlab using vector operations. We tested the alternative choice of $H^k = I$, which does not require Hessian evaluation, but its overall performance was worse. If Hessian computation is expensive, a compromise would be to recompute the Hessian diagonal once every few iterations. We choose the index subset $\mathcal{J}^k$ by either (i) the restricted Gauss-Seidel rule (2.7), whereby $\mathcal{N}$ is partitioned into $n_b \in \{5, 10, n\}$

subsets with $\lceil n/n_b \rceil$ elements each (except possibly the last subset which may have fewer elements) and $\mathcal{J}^k$ cycles through these subsets, or (ii) the Gauss-Southwell-r rule (2.8) with $D^k = H^k$,

$$\mathcal{J}^k = \left\{ j \mid |d_{D^k}(x^k; j)| \geq \upsilon^k \|d_{D^k}(x^k; \mathcal{N})\|_\infty \right\},$$
$$\upsilon^{k+1} = \begin{cases} \max\{10^{-4}, \upsilon^k/10\} & \text{if } \alpha^k > 10^{-3} \\ \min\{.9, 50\upsilon^k\} & \text{if } \alpha^k < 10^{-6} \\ \upsilon^k & \text{else} \end{cases}$$

(initially $\upsilon^0 = .5$) or (iii) the Gauss-Southwell-q rule (2.10) with $D^k = H^k$,

$$\mathcal{J}^k = \left\{ j \mid q_{D^k}(x^k; j) \leq \upsilon^k \min_i q_{D^k}(x^k; i) \right\},$$
$$\upsilon^{k+1} = \begin{cases} \max\{10^{-4}, \upsilon^k/10\} & \text{if } \alpha^k > 10^{-3} \\ \min\{.9, 50\upsilon^k\} & \text{if } \alpha^k < 10^{-6} \\ \upsilon^k & \text{else} \end{cases}$$

(initially $\upsilon^0 = .5$). The above updating formulas for υ^k in (ii) and (iii) are guided by the observation that smaller υ^k results in more coordinates being updated but a smaller stepsize α^k, while a larger υ^k has the opposite effect. Thus if α^k is large, we decrease υ^k and if α^k is small, we increase υ^k. The thresholds 10^{-3} and 10^{-6} were found after some experimentation to work well on our test problems. The stepsize α^k is chosen by the Armijo rule (2.4) with

$$\sigma = .1, \quad \beta = .5, \quad \theta = 0, \qquad \alpha^0_{\text{init}} = 1, \quad \alpha^k_{\text{init}} = \min\left\{ \frac{\alpha^{k-1}}{\beta}, 1 \right\} \ \forall k \geq 1.$$

We experimented with other values of $0 \leq \theta < 1$, but the cpu times and the number of iterations did not change appreciably in our tests.

Each CGD iteration requires 1 gradient evaluation and 1 Hessian diagonal evaluation to find the direction d^k, and at least 1 function evaluation to find the stepsize α^k. These are the dominant computations. For the CUTEr test functions, Hessian evaluation is the most dominant computation when the Hessian is dense. (CUTEr does not offer the option of evaluating only the Hessian diagonals.)

Since H^k is diagonal, the CGD method resembles a block coordinate version of a diagonally scaled steepest descent method [6, page 71] when $c = 0$. As such, the convergence rate of the method is likely slow when the Hessian $\nabla^2 f(x^k)$ is far from being diagonally dominant, as was observed on some of the functions from Table 2.1, such as LR1, LR1Z, and VD. This motivated us to introduce two techniques to accelerate the convergence, which we describe below.

The first technique uses an active-set identification strategy of Facchinei, Fischer, and Kanzow [24] (also see [25, Section 6.7]) to estimate which components of x would be nonzero at a solution and then uses a fast method for unconstrained smooth optimization to update these components. The method we chose is the limited-memory BFGS (L-BFGS) method of Nocedal [76, 77]. In particular, we store the m ($m \geq 1$) most recent pairs of Δx and Δg that make sufficiently acute angles. More precisely, we store $\Delta x^k = x^k - x^{k-1}$ and $\Delta g^k = g^k - g^{k-1}$ (with $g^k = \nabla f(x^k)$) whenever

$$\|\Delta g^k\| > 10^{-20}, \qquad \frac{\Delta x^{k^T} \Delta g^k}{\|\Delta g^k\|^2} > \frac{10^{-10}}{\max_j H^k_{jj}}.$$

In an acceleration step at x^k, we use the L-BFGS formula (with $m = 5$) to construct a positive definite Hessian inverse approximation B^k and set

$$d^k_{\mathcal{J}^k} = -B^k_{\mathcal{J}^k \mathcal{J}^k} \nabla_{x_{\mathcal{J}^k}} F_c(x^k), \qquad d^k_j = 0 \; \forall j \notin \mathcal{J}^k,$$

where $\mathcal{J}^k = \left\{ j \mid |x^k_j| > \rho\left(\|d_{H^k}(x^k; \mathcal{N})\|_\infty \right) \right\}$ with the identification function $\rho(t) = \dfrac{-.0001}{\ln(\min\{.1, .01t\})}$. We then update $x^{k+1} = x^k + \alpha^k d^k$ with α^k chosen by the Armijo rule with $\sigma = .1$, $\beta = .5$, $\theta = 0$, and $\alpha^k_{\text{init}} = 1$. This acceleration step is invoked at iteration k whenever $k \geq 10$ and $k < 50 \pmod{100}$. We choose 50 since L-BFGS typically terminates in less than 50 iterations on the test functions when $c = 0$.

The second technique is motivated by the rank-1 Hessian for the functions LR1 and LR1Z. In an acceleration step at x^k, we choose h^k to satisfy the rank-1 secant equation

$$(h^k h^{k^T}) s^k = y^k,$$

where s^k and y^k are the most recently stored pair of Δx and Δg. This yields $h^k = y^k/\sqrt{s^{kT}y^k}$. We next solve the subproblem with rank-1 Hessian

$$\min_d \; g^{kT}d + \frac{1}{2}(h^{kT}d)^2 + c\|x^k + d\|_1.$$

This subproblem need not have an optimal solution (e.g., when $h^k_j = 0$ and $|g^k_j| > c$ for some j), but if it has an optimal solution, then there exists an optimal solution d^k with at most one nonzero component, which can be computed efficiently using Matlab's vector operations. (In general, if the subproblem with rank-p Hessian has an optimal solution, then there exists an optimal solution with at most p nonzero components.) We then update $x^{k+1} = x^k + \alpha^k d^k$ with α^k chosen as in the L-BFGS acceleration step. This second acceleration step is invoked once every 10 consecutive CGD iterations.

We terminate the CGD method when

$$\|H^k d_{H^k}(x^k; \mathcal{N})\|_\infty \leq 10^{-4}. \tag{2.58}$$

Here we scale $d_{H^k}(x^k; \mathcal{N})$ by H^k to reduce its sensitivity to H^k. We can alternatively use the criterion $\|d_I(x^k; \mathcal{N})\|_\infty \leq 10^{-4}$. The advantage of (2.58) is that $d_{H^k}(x^k; \mathcal{N})$ is already computed by the CGD method, unlike $d_I(x^k; \mathcal{N})$. In a few cases where $\nabla^2 f$ is ill-conditioned, the Armijo descent condition (2.4) eventually cannot be satisfied by any $\alpha^k > 0$ due to cancellation error in the function evaluations. (In Matlab Version 7.0, floating point subtraction is accurate up to 15 digits only.) In these cases, no further progress is possible so we exit the method when (2.4) remains unsatisfied after α^k reaches 10^{-30}.

2.6.3 L-BFGS-B and MINOS

The ℓ_1-regularized problem (1.3) can be formulated as a bound-constrained smooth optimization problem:

$$\min_{y \geq 0, z \geq 0} \; f(y - z) + c\, e^T(y + z),$$

where e is the vector of 1s, to which many methods can be applied for its solution. Thus, it is of interest to compare the CGD method with such methods. We considered two such methods. One is L-BFGS-B, a Fortran implementation of a limited memory algorithm for large-scale bound-constrained smooth optimization [109]. The public code was downloaded from http://www.ece.northwestern.edu/~nocedal/lbfgsb.html. A second is MINOS (Version 5.5.1), which has a Fortran implementation of an active-set method for linearly constrained smooth optimization [75]. To accommodate problems with $n = 1000$, we set Superbasics limit to $2n + 1$ and Workspace to 5,000,000 in MINOS. The objective function and its gradient are coded in Fortran, with f taken from Table 2.1. For a given starting point x^0 for (1.3), we accordingly initialize $y^0 = \max\{x^0, 0\}$ and $z^0 = \max\{-x^0, 0\}$, with the "max" taken componentwise.

2.6.4 Numerical Results

We now report the performance of the CGD method using either the restricted Gauss-Seidel rule or the Gauss-Southwell-r (GS-r) rule or the Gauss-Southwell-q (GS-q) rule, with or without the aforementioned acceleration techniques, and we compare it with the performances of L-BFGS-B and MINOS. All runs are performed on an HP DL360 workstation, running Red Hat Linux 3.5 and Matlab (Version 7.0). All Fortran codes are compiled using the Gnu F-77 compiler (Version 3.2.57). Tables 2.3–2.7[3] show the final objective value, the cpu time (in seconds), and the number of nonzero components ($\sharp$nz) in the final solution found. (A component is considered to be nonzero if its absolute value exceeds 10^{-15}.) For each function, three different values of c are chosen to track changes in the solution sparsity $\sharp$nz. In Tables 2.4–2.6, different starting points are used. In Tables 2.4–2.7, the number of L-BFGS

[3] [a]: CGD exited due to the Armijo stepsize in an CGD iteration reaching 10^{-30}.
[b]: CGD exited due to the Armijo stepsize in an L-BFGS acceleration step reaching 10^{-30}.
[c]: L-BFGS-B exited due to the objective value cannot be improved upon.
[d]: MINOS exited due to the current point cannot be improved upon.
[e]: MINOS exited due to the problem being badly scaled.
[f]: CGD is terminated using tolerance 10^{-7}.

Table 2.3: Comparing the CGD method using the Gauss-Seidel rule and the Gauss-Southwell rules, without acceleration steps, on the test functions from Table 2.1, with x^0 given as in [73].

Name	c	CGD-GSeidel ($n_b = n$)	CGD-GSeidel ($n_b = 5$)	CGD-GS-r	CGD-GS-q
		♯nz/obj/cpu	♯nz/obj/cpu	♯nz/obj/cpu	♯nz/obj/cpu
BAL	1	[a]1000/249755/.7	>5h	[a]1000/1000.00/.1	[a]1000/1000.00/.1
	10	[a]1000/259247/.04	>5h	[a]1000/9999.98/.1	[a]1000/9999.98/.2
	100	[a]1000/344302/6.3	>5h	>5h	>5h
BT	.1	1000/70.3320/40.0	1000/70.3320/.8	1000/70.3320/.1	1000/70.3320/.1
	1	1000/671.819/48.2	1000/671.819/1.0	1000/671.819/.2	1000/671.819/.2
	10	0/1000.00/6.4	0/1000.00/.07	0/1000.00/.02	0/1000.00/.02
DBV	.1	3/0.00000/20.3	0/0.00000/.9	2/0.00000/.04	2/0.00000/.04
	1	0/0.00000/3.0	0/0.00000/.03	2/0.00000/.01	2/0.00000/.02
	10	0/0.00000/3.0	0/0.00000/.03	0/0.00000/.01	0/0.00000/.02
ER	1	1000/436.250/1642.3	1000/436.250/4.9	1000/436.250/.8	1000/436.250/.8
	10	0/500.000/28.3	0/500.000/.4	0/500.000/.1	0/500.000/.1
	100	0/500.000/5.9	0/500.000/.07	0/500.000/.01	0/500.000/.01
TRIG	.1	0/0.00000/131.7	0/0.00000/.6	0/0.00000/.1	0/0.00000/.1
	1	0/0.00000/8.8	0/0.00000/.08	0/0.00000/.02	0/0.00000/.02
	10	0/0.00000/2.3	0/0.00000/.02	0/0.00000/.01	0/0.00000/.01
EPS	1	1000/351.146/194.6	1000/351.146/1.4	1000/351.146/.3	1000/351.146/.3
	10	250/1250.00/20.8	200/1250.00/.2	250/1250.00/.03	250/1250.00/.04
	100	0/1250.00/6.1	0/1250.00/.07	0/1250.00/.01	0/1250.00/.01
LR1	.1	[a]1000/50399.4/.1	>5h	>5h	>5h
	1	[a]1000/501748/.1	>5h	>5h	>5h
	10	[a]1000/5015230/.1	>5h	>5h	>5h
LR1Z	.1	[a]1000/44894.4/.1	>5h	>5h	>5h
	1	[a]1000/446684/.1	>5h	>5h	>5h
	10	[a]1000/4464582/.1	>5h	>5h	>5h
LFR	.1	1000/98.5000/.9	[a]1000/98.5000/.04	1000/98.5000/.01	1000/98.5000/.01
	1	1000/751.000/.9	1000/751.000/.02	1000/751.000/.01	1000/751.000/.01
	10	0/1001.00/.9	0/1001.00/.02	0/1001.00/.01	0/1001.00/.01
VD	1	$999/3.51{\cdot}10^{11}$/.1	>5h	>5h	>5h
	10	$999/3.51{\cdot}10^{11}$/.1	>5h	>5h	>5h
	100	$999/3.52{\cdot}10^{11}$/.1	>5h	>5h	>5h

acceleration steps and rank-1 acceleration steps are also shown. In our experience, CGD-GS-r and CGD-GS-q have comparable performances. Also, we found CGD-GSeidel to have better performance with $n_b = 5$ than with $n_b = 10$ or $n_b = n$.

From Table 2.3, we see that CGD-GS-r and CGD-GS-q are typically faster than CGD-GSeidel. But CGD-GS-r and CGD-GS-q are still too slow (more than 5 hours of cpu time) on functions whose Hessian are far from being diagonally dominant, like BAL, LR1, LR1Z, and VD. From Table 2.4, we see that the acceleration steps improve the performance of CGD-GS-r and CGD-GS-q significantly on these functions. We also tested CGD-GSeidel with acceleration steps, but its performance is not better

Table 2.4: Comparing the CGD method using the Gauss-Southwell rules, with or without acceleration steps, on test functions from Table 2.1, with x^0 given as in [73].

Name	c	CGD-GS-r	CGD-GS-r-acc	CGD-GS-q	CGD-GS-q-acc
		♯nz/obj/cpu(iter)	♯nz/obj/cpu (CGD/L-BFGS/R1)	♯nz/obj/cpu(iter)	♯nz/obj/cpu (CGD/L-BFGS/R1)
BAL	1	[a]1000/1000.00/.1(12)	1000/1000.00/.1(10/22/1)	[a]1000/1000.00/.1(20)	1000/1000.00/.2(10/29/1)
	10	[a]1000/9999.98/.1(12)	[b]1000/9999.97/.1(10/16/1)	[a]1000/9999.98/.2(56)	1000/9999.97/.1(10/21/1)
	100	>5h	1000/99997.5/.1(10/9/1)	>5h	[b]1000/99997.5/.1(10/18/1)
BT	.1	1000/70.3320/.1(55)	1000/70.3320/.1(10/15/1)	1000/70.3320/.1(55)	1000/70.3320/.1(10/14/1)
	1	1000/671.819/.2(71)	1000/671.819/.1(10/19/1)	1000/671.819/.2(71)	1000/671.819/.1(10/19/1)
	10	0/1000.00/.02(6)	0/1000.00/.03(8/0/1)	0/1000.00/.02(6)	0/1000.00/.03(8/0/1)
DBV	.1	2/0.00000/.04(10)	0/0.00000/.02(2/0/1)	2/0.00000/.04(10)	0/0.00000/.02(2/0/1)
	1	2/0.00000/.01(3)	0/0.00000/.02(2/0/1)	2/0.00000/.02(3)	0/0.00000/.02(2/0/1)
	10	0/0.00000/.01(3)	0/0.00000/.02(2/0/1)	0/0.00000/.02(3)	0/0.00000/.02(2/0/1)
ER	1	1000/436.250/.8(346)	1000/436.250/.3(11/40/1)	1000/436.250/.8(309)	1000/436.250/.2(10/38/1)
	10	0/500.000/.1(32)	0/500.000/.3(11/38/1)	0/500.000/.1(28)	0/500.000/.3(11/40/1)
	100	0/500.000/.01(5)	0/500.000/.03(8/0/1)	0/500.000/.01(5)	0/500.000/.04(8/0/1)
TRIG	.1	0/0.00000/.1(42)	1000/0.00028/.1(11/10/0)	0/0.00000/.1(42)	0/0.00000/.1(12/9/0)
	1	0/0.00000/.02(5)	0/0.00000/.02(5/0/0)	0/0.00000/.02(6)	0/0.00000/.02(6/0/0)
	10	0/0.00000/.01(1)	0/0.00000/.01(1/0/0)	0/0.00000/.01(1)	0/0.00000/.01(1/0/0)
EPS	1	1000/351.146/.3(72)	1000/351.146/.3(10/37/1)	1000/351.146/.3(71)	1000/351.146/.2(10/30/1)
	10	250/1250.00/.03(10)	249/1250.00/.1(10/0/1)	250/1250.00/.04(10)	250/1250.00/.05(10/0/1)
	100	0/1250.00/.01(3)	0/1250.00/.01(2/0/1)	0/1250.00/.01(3)	0/1250.00/.02(2/0/1)
LR1	.1	>5h	1/249.625/.1(10/0/2)	>5h	1/249.625/.1(10/0/2)
	1	>5h	1/249.625/.1(10/0/1)	>5h	1/249.625/.1(10/0/2)
	10	>5h	1/249.625/.1(10/0/2)	>5h	1/249.625/.05(8/0/1)
LR1Z	.1	>5h	1/251.125/.1(10/0/2)	>5h	1/251.125/.1(10/0/2)
	1	>5h	1/251.125/.1(10/0/1)	>5h	1/251.125/.1(10/0/1)
	10	>5h	1/251.125/.1(10/0/2)	>5h	1/251.125/.1(10/0/1)
LFR	.1	1000/98.5000/.01(1)	1000/98.5000/.01(1/0/0)	1000/98.5000/.01(1)	1000/98.5000/.01(1/0/0)
	1	1000/751.000/.01(1)	1000/751.000/.01(1/0/0)	1000/751.000/.01(1)	1000/751.000/.01(1/0/0)
	10	0/1001.00/.01(1)	0/1001.00/.01(1/0/0)	0/1001.00/.01(1)	0/1001.00/.01(1/0/0)
VD	1	>5h	1000/937.594/1.7 (191/240/21)	>5h	1000/937.594/.6 (56/80/5)
	10	>5h	[b]1000/6726.81/64.6 (5635/6247/626)	>5h	[b]1000/6726.81/42.6 (3791/4199/420)
	100	>5h	[b]999/55043.1/51.8 (4600/5106/511)	>5h	[b]1000/55043.1/106.2 (8291/9198/920)

than CGD-GS-q-acc and so we do not report it here.

From Tables 2.5 and 2.6, we see that CGD-GS-r-acc and CGD-GS-q-acc are competitive with MINOS in terms of solution accuracy (as measured by the final objective value), and are generally faster in terms of cpu time (except on VD). L-BFGS-B is fast, but often exits when still far from a solution with a large projected gradient. This is due to the relative improvement in objective value being below $factr \cdot epsmch$, where $factr = 10^7$ and $epsmch$ is the machine precision generated by the code (about

Table 2.5: Comparing the CGD method using the Gauss-Southwell rules and acceleration steps with L-BFGS-B and MINOS on test functions from Table 2.1, with $x^0 = (1, 1, ..., 1)^T$.

Name	c	L-BFGS-B	MINOS	CGD-GS-r-acc	CGD-GS-q-acc
		♯nz/obj/cpu	♯nz/obj/cpu	♯nz/obj/cpu (CGD/L-BFGS/R1)	♯nz/obj/cpu (CGD/L-BFGS/R1)
BAL	1	[c]1000/1000.00/.02	1000/1000.00/49.9	1000/1000.00/.1(10/17/1)	1000/1000.00/.1(10/10/1)
	10	[c]1000/9999.98/.03	1000/9999.97/48.4	1000/9999.97/.1(10/14/1)	1000/9999.98/.2(10/9/1)
	100	[c]1000/99997.5/.1	1000/99997.5/48.9	[b]1000/99997.5/.1(10/18/1)	[b]1000/99997.5/.1(10/15/1)
BT	.1	[c]1000/84.0033/.02	1000/71.725/100.6	1000/72.2619/.9(109/117/4)	1000/71.7481/.9(111/97/0)
	1	[c]981/668.724/.2	997/672.418/94.7	1000/626.670/41.8 (4219/4267/42)	1000/626.670/42.4 (4156/4154/5)
	10	0/1000.00/.00	0/1000.00/56.0	0/1000/.01(1/0/0)	0/1000.00/.01(1/0/0)
DBV	.1	[c]999/83.4557/.01	0/0.00000/51.5	0/0.00000/.5(11/40/2)	0/0.00000/.5(11/40/2)
	1	0/0.00000/.01	0/0.00000/50.8	0/0.00000/.03(5/0/1)	2/0.00000/.03(3/0/1)
	10	0/0.00000/.00	0/0.00000/52.5	0/0.00000/.01(1/0/0)	0/0.00000/.01(1/0/0)
ER	1	1000/436.250/.1	1000/436.250/71.5	1000/436.250/.2(10/38/1)	1000/436.250/.1(10/24/1)
	10	[c]500/1721.15/.00	0/500.000/50.2	449/500.006/.3(11/40/1)	0/500.000/.3(11/40/1)
	100	0/500.000/.00	0/500.000/52.4	0/500.000/.03(7/0/1)	0/500.000/.03(7/0/1)
TRIG	.1	[c]1000/14.1282/.1	0/0.00000/58.5	6/3.13589/.6(55/45/6)	1/.626211/.6(29/40/4)
	1	0/0.00000/.1	1/6.21995/62.0	6/31.2477/.7(55/47/6)	1/6.21364/.5(47/40/6)
	10	0/0.00000/.1	0/0.00000/61.9	1/187.021/.6(47/40/6)	1/61.2209/.5(38/40/5)
EPS	1	[c]999/352.526/.05	1000/351.146/60.3	1000/351.146/.3(10/40/1)	1000/351.146/.3(13/40/2)
	10	1/1250.00/.01	243/1250.00/44.2	250/1250.00/.1(9/0/1)	249/1250.00/.1(8/0/1)
	100	0/1250.00/.01	0/1250.00/51.5	0/1250.00/.01(1/0/0)	0/1250.00/.01(2/0/1)
LR1	.1	[c]1000/424.663/.00	[d]2/249.625/59.7	1/249.625/.1(10/0/2)	1/249.625/.1(10/0/2)
	1	[c]1000/2000.00/.01	[d]1/249.625/57.2	1/249.625/.1(10/0/1)	1/249.625/.1(10/0/2)
	10	[c]1000/17753.4/.01	1/249.625/58.0	1/249.625/.1(10/0/2)	1/249.625/.05(8/0/1)
LR1Z	.1	[c]1000/426.087/.00	[d]4/251.125/59.2	1/251.125/.1(10/0/2)	1/251.125/.1(10/0/2)
	1	[c]1000/2000.75/.01	[d]3/251.125/58.4	1/251.125/.1(10/0/1)	1/251.125/.1(10/0/1)
	10	[c]1000/17747.3/.00	1/251.125/59.7	1/251.125/.1(10/0/2)	1/251.125/.1(10/0/1)
LFR	.1	1000/98.5000/.00	1000/98.5000/77.2	1000/98.5000/.01(1/0/0)	1000/98.5000/.01(1/0/0)
	1	1000/751.000/.01	1000/751.000/73.8	1000/751.000/.01(1/0/0)	1000/751.000/.01(1/0/0)
	10	0/1001.00/.00	0/1001.00/53.3	0/1001.00/.01(1/0/0)	0/1001.00/.01(1/0/0)
VD	1	[c]1000/1000.00/.00	1000/937.594/43.0	1000/937.594/.9 (100/139/11)	1000/937.594/.5 (55/59/6)
	10	[c]$974/5.18\cdot10^{12}$/2.3	413/6726.81/56.9	[b]1000/6726.81/59.9 (5230/5803/581)	[b]1000/6726.81/60.3 (5140/5698/571)
	100	[c]996/75135.5/.2	136/55043.1/57.4	[b]1000/55043.1/83.3 (6850/7604/761)	[b]1000/55043.1/88.1 (7030/7804/781)

10^{-19} in our tests). We experimented with $factr$ set to zero but it did not change significantly the results.

Thus MINOS seems more robust than L-BFGS-B, though it is slower (possibly due to the many active bounds at a solution). For the nonconvex functions BT and TRIG, multiple local minima exist and, depending on the starting point, the methods can converge to different local minima with different objective value.

Table 2.7 reports the performance of CGD-GS-r-acc and CGD-GS-q-acc on the

Table 2.6: Comparing the CGD method using the Gauss-Southwell rules and acceleration steps with L-BFGS-B and MINOS on test functions from Table 2.1, with $x^0 = (-1, -1, ..., -1)^T$.

Name	c	L-BFGS-B	MINOS	CGD-GS-r-acc	CGD-GS-q-acc
		♯nz/obj/cpu	♯nz/obj/cpu	♯nz/obj/cpu (CGD/L-BFGS/R1)	♯nz/obj/cpu (CGD/L-BFGS/R1)
BAL	1	[c]1000/1000.00/.1	1000/1000.00/39.9	1000/1000.00/.2(10/29/1)	1000/1000.00/.2(10/27/1)
	10	[c]1000/9999.97/.2	1000/9999.97/50.0	[b]1000/9999.97/.1(10/25/1)	1000/9999.97/.2(10/29/1)
	100	[c]1000/99997.5/.2	1000/99997.5/50.6	1000/99997.5/.1(10/26/1)	1000/99997.5/.2(10/34/1)
BT	.1	[c]1000/70.9405/.1	1000/70.3320/99.0	1000/70.3320/.1(10/15/1)	1000/70.3320/.1(10/14/1)
	1	999/671.773/.1	999/671.773/101.1	1000/671.819/.1(10/19/1)	1000/671.819/.1(10/19/1)
	10	0/1000.00/.01	0/1000.00/77.1	0/1000.00/.03(8/0/1)	0/1000.00/.03(8/0/1)
DBV	.1	[c]1000/82.7786/.01	0/0.00000/66.0	0/0.00000/.4(11/40/2)	0/0.00000/.4(11/40/2)
	1	[c]4/6.47238/.01	0/0.00000/65.8	0/0.00000/.03(5/0/1)	0/0.00000/.02(3/0/1)
	10	0/0.00000/.01	0/0.00000/66.1	0/0.00000/.01(1/0/0)	0/0.00000/.01(1/0/0)
ER	1	1000/436.250/.04	1000/436.250/86.9	1000/436.250/.2(10/33/1)	1000/436.250/.1(11/22/0)
	10	0/500.000/.03	0/500.000/74.2	0/500.000/.3(11/40/1)	1000/500.024/.2(15/34/0)
	100	0/500.000/.01	0/500.000/62.1	0/500.000/.03(7/0/1)	0/500.000/.01(2/0/0)
TRIG	.1	[c]1000/12.7569/.1	0/0.00000/104.4	1/1.25435/.6(55/44/6)	1/71.6259/.6(55/48/6)
	1	[c]1000/181.247/.1	0/0.00000/106.8	3/50.1248/.7(56/48/6)	1/364.351/.6(55/44/6)
	10	[c]1000/2818.55/.1	0/0.00000/110.1	1/124.051/.6(47/40/6)	1/1820.88/.5(55/42/6)
EPS	1	1000/351.146/.1	1000/351.146/106.7	1000/351.146/.2(10/22/1)	1000/351.146/.2(10/26/1)
	10	249/1250.00/.01	0/1250.00/74.6	250/1250.00/.1(9/0/1)	250/1250.00/.1(8/0/1)
	100	0/1250.00/.00	0/1250.00/63.4	0/1250.00/.01(1/0/0)	0/1250.00/.02(2/0/1)
LR1	.1	[c]1000/424.663/.00	[d]4/249.625/85.0	1/249.625/.1(10/0/1)	1/249.625/.1(10/0/2)
	1	[c]1000/2000.00/.01	[d]1/249.625/85.4	1/249.625/.1(10/0/2)	1/249.625/.1(10/0/1)
	10	[c]1000/17753.4/.00	1/249.625/82.0	1/249.625/.1(10/0/1)	1/249.625/.1(10/0/2)
LR1Z	.1	[c]1000/426.087/.00	[d]2/251.125/84.3	1/251.125/.1(10/0/2)	1/251.125/.1(10/0/1)
	1	[c]1000/2000.75/.00	1/251.125/85.3	1/251.125/.1(10/0/1)	1/251.125/.1(10/0/1)
	10	[c]1000/17747.3/.01	1/251.125/84.8	1/251.125/.1(10/0/1)	1/251.125/.1(9/0/1)
LFR	.1	1000/98.5000/.01	1000/98.5000/57.4	1000/98.5000/.01(1/0/0)	1000/98.5000/.01(1/0/0)
	1	1000/751.000/.00	1000/751.000/59.2	1000/751.000/.01(1/0/0)	1000/751.000/.01(1/0/0)
	10	0/1001.00/.00	0/1001.00/66.4	0/1001.00/.01(1/0/0)	0/1001.00/.01(1/0/0)
VD	1	[c]1000/1836.78/.3	[e]999/100401e+24/.1	1000/937.594/2.6 (235/271/26)	1000/937.594/.6(56/77/5) (56/77/5)
	10	[c]1000/25653.0/.2	[e]999/100401e+24/.1	1000/6726.81/27.7 (2665/2954/296)	[b]1000/6726.81/29.6 (2711/3002/300)
	100	[c]1000/248974/.2	[e]999/100401e+24/.1	[b]999/55043.1/50.9 (4600/5135/511)	[b]1000/55043.1/105.0 (8156/9052/905)

CUTEr test functions from Table 2.2. Both are able to meet the termination criterion (2.58) in typically under a second, except on NONCVXU2 and PENALTY1. On PENALYT1, the termination tolerance 10^{-4} in (2.58) was too loose, with the final objective value accurate up to only 1 or 2 significant digits, so we tightened it to 10^{-9}. The final objective value for other functions appear to be accurate up to 5 significant digits, as tightening the tolerance to 10^{-6} did not change them. Notice that, on INDEF, LIARWHD, NONCVXU2, PENALTY1, WOODS, for which f is nonconvex,

Table 2.7: Comparing the CGD method using the Gauss-Southwell rules and acceleration steps on CUTEr test functions from Table 2.2, with x^0 as given.

Name	c	CGD-GS-r-acc	CGD-GS-q-acc
		♯nz/obj/cpu(CGD/L-BFGS/R1)	♯nz/obj/cpu(CGD/L-BFGS/R1)
EG2	.1	1/-998.890/.02(2/0/1)	1/-998.890/.02(2/0/1)
	1	1/-998.377/.02(2/0/1)	1/-998.377/.02(2/0/1)
	10	1/-993.290/.02(2/0/1)	1/-993.290/.02(2/0/1)
EXTROSNB	.1	5/.235809/.8(61/42/2)	5/.235809/.8(59/50/2)
	1	3/.873442/.2(14/13/1)	2/.873441/.8(59/48/2)
	10	0/1.00000/.04(4/0/1)	0/1.00000/.5(12/40/1)
INDEF	1	[b]1000/-499.000/1.5(58/41/3)	1000/-499.000/.9(30/40/1)
	10	2/-301.161/.2(10/5/1)	2/-18.4175/.2(11/5/0)
	100	2/-197.836/.2(10/4/1)	3/499.605/.1(5/0/0)
LIARWHD	.1	1000/101.025/.2(10/19/1)	1000/97.5328/.1(10/8/1)
	1	1000/750.203/.3(10/26/1)	1000/750.203/.1(10/5/1)
	10	0/1000.00/.5(11/40/2)	0/1000.00/.04(4/0/1)
NONCVXU2	.1	948/2390.60/7.0(375/440/40)	957/2710.90/12.3(625/690/40)
	1	683/3120.28/13.2(687/712/24)	677/3124.66/8.7(451/452/10)
	10	0/4000.00/1.9(91/90/9)	5/4000.00/1.9(92/90/8)
PENALTY1	.01	[f]1/.0149673/37.7(11/15/1)	[f]1/.0149673/88.8(25/40/2)
	.1	[f]1/.0571739/14.9(10/0/1)	[f]0/.072500/14.9(10/0/0)
	1	[f]0/.072500/12.1(8/0/1)	[f]0/.072500/14.7(10/0/0)
WOODS	1	1000/985.710/2.1(149/160/12)	1000/985.710/2.1(149/157/12)
	10	750/8655.68/.8(59/56/2)	1000/8655.70/.2(11/25/0)
	100	249/10500.0/.5(11/40/1)	750/10500.7/.5(12/40/0)
QUARTC	.1	1000/50028.1/.2(11/18/0)	1000/50028.1/.2(11/25/0)
	1	1000/500028/.1(11/15/0)	1000/500028/.2(11/22/0)
	10	$999/4.99482\cdot10^6$/.1(11/13/0)	$1000/4.99482\cdot10^6$/.2(11/26/0)
DIXON3DQ	.1	6/.470417/.6(52/40/0)	6/.470417/.6(46/40/0)
	1	2/1.62500/.02(3/0/1)	2/1.62500/.05(7/0/0)
	10	0/2.00000/.01(1/0/0)	0/2.00000/.02(4/0/0)
TRIDIA	.1	8/.185656/.5(51/40/6)	8/.185656/.6(58/48/3)
	1	2/.911765/.5(40/40/3)	2/.911765/.5(43/40/2)
	10	0/1.00000/.3(11/40/2)	0/1.00000/.3(12/40/1)

CGD-GS-r and CGD-GS-q can terminate at different solutions, depending on the starting point x^0.

2.7 Conclusions and Extensions

We have presented a block coordinate gradient descent method for minimizing the sum of a smooth function and a convex separable function. The method may be viewed as a hybrid of gradient-projection and coordinate descent methods, or as a block coordinate version of descent methods in [9, 36]. We analyzed the global convergence and asymptotic convergence rate of the method. We also presented numerical results

to verify the practical efficiency of the method.

We can relax the Armijo descent condition (2.4) by replacing Δ^k with its upper bound $(\theta-1){d^k}^T H^k d^k$ (see (2.20)), i.e.,

$$F_c(x^k+\alpha^k d^k) \le F_c(x^k) + \alpha^k \sigma(\theta-1){d^k}^T H^k d^k. \tag{2.59}$$

The global convergence analysis in Theorem 2.1 (except (d)) can be extended accordingly. The convergence rate analysis in Theorem 2.2 can be similarly extended, provided that $\alpha^k = 1$ for all k sufficiently large (so that the last term in (2.34) equals zero). Using Lemma 2.1 and the fact that, under assumption (2.16), $f(x+d)-f(x) \le \nabla f(x)^T d + L\|d\|^2/2$ for all $x, x+d \in \mathrm{dom}P$ (see [6, page 667] or the proof of Lemma 2.5(b)), it is readily seen that the latter holds if we choose $\alpha^k_{\text{init}} = 1$ and $\theta \ge L/(2\underline{\lambda})$. A similar convergence rate result was shown by Fukushima and Mine for their method [36, Theorem 5.1] under the additional assumption that f is (locally) strongly convex. On the other hand, Theorem 2.4 does not seem amenable to a similar extension, due to the presence of an additional term $-q_{D^k}(x^k;(\mathcal{J}^k)^c)$ in (2.47), which is in the order of $-\Delta^k$; see (2.52) and (2.53). If the Lipschitz constant L is unknown, we can still ensure that $\alpha^k = 1$ by adaptively scaling H^k when generating d^k, analogous to the Armijo rule along the projection arc for constrained smooth optimization [6, page 236]. In particular, we choose s^k to be the largest element of $\{s\beta^j\}_{j=0,1,\ldots}$ $(s>0)$ such that

$$d^k = d_{H^k/s^k}(x^k;\mathcal{J}^k)$$

satisfies the relaxed Armijo descent condition (2.59) with $\alpha^k = 1$. This adaptive scaling strategy is more expensive computationally since d^k needs to be recomputed each time s^k is changed. Still, if P is separable and we choose H^k to be diagonal, then d^k is relatively cheap to recompute.

There are many directions for future research. For example, in our current implementation of the CGD method, we used diagonal H^k. How about block-diagonal H^k? (For efficiency, this may need to be coded in Fortran since Matlab's vector operations

might not be usable.) Can other acceleration techniques be developed? How would the CGD method perform on bound-constrained problems? Can the assumption on P in Theorem 2.1(d) be dropped? Can a linear convergence rate result similar to Theorem 2.4 be proved when $\{\mathcal{J}^k\}$ is chosen by the Gauss-Southwell-r rule?

Chapter 3

A (BLOCK) COORDINATE GRADIENT DESCENT METHOD FOR LINEARLY CONSTRAINED SMOOTH OPTIMIZATION AND SUPPORT VECTOR MACHINES TRAINING

In this chapter, we study the CGD method for solving (1.6), in particular (1.7). Our method is closely related to decomposition methods currently popular for SVM training. We describe the CGD method formally and show the global convergence and asymptotic convergence rate of the method when the coordinate block is chosen by a Gauss-Southwell-type rule. We show that, for SVM QP with n variables, this rule can be implemented in $O(n)$ operations using Rockafellar's notion of conformal realization. Thus, for SVM training, our method requires only $O(n)$ operations per iteration and, in contrast to existing decomposition methods, achieves linear convergence without additional assumptions. We report our numerical experience with the method on some large SVM QP arising from two-class data classification. Our experience suggests that the method can be efficient for SVM training with nonlinear kernel. This chapter is based on the paper [104] co-authored with P. Tseng.

3.1 (Block) Coordinate Gradient Descent Method

In our method, we use $\nabla f(x)$ to build a quadratic approximation of f at x and apply coordinate descent to generate an improving feasible direction d at x. More precisely, we choose a nonempty subset $\mathcal{J} \subseteq \mathcal{N}$ and a symmetric matrix $H \in \mathbb{R}^{n\times n}$ (approximating the Hessian $\nabla^2 f(x)$), and move x along the direction $d = d_H(x; \mathcal{J})$,

where

$$d_H(x;\mathcal{J}) \stackrel{\text{def}}{=} \arg\min_{d\in\mathbb{R}^n}\left\{ \nabla f(x)^T d + \frac{1}{2}d^T H d \;\mid\; x+d\in X,\; d_j = 0\ \forall j \notin \mathcal{J}\right\}. \quad (3.1)$$

Here $d_H(x;\mathcal{J})$ depends on H through $H_{\mathcal{J}\mathcal{J}}$ only.

To ensure that $d_H(x;\mathcal{J})$ is well defined, we assume that $H_{\mathcal{J}\mathcal{J}}$ is positive definite on Null$(A_{\mathcal{J}})$ or, equivalently, $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succ 0_n$, where $A_{\mathcal{J}}$ denotes the submatrix of A comprising columns indexed by $\mathcal{J}$ and $B_{\mathcal{J}}$ is a matrix whose columns form an orthonormal basis for Null$(A_{\mathcal{J}})$. For (1.7), we can choose H such that $H_{\mathcal{J}\mathcal{J}} = Q_{\mathcal{J}\mathcal{J}}$ if $B_{\mathcal{J}}^T Q_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succ 0_n$ and otherwise $H_{\mathcal{J}\mathcal{J}} = Q_{\mathcal{J}\mathcal{J}} + \rho I$ with $\rho > 0$ such that $B_{\mathcal{J}}^T Q_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} + \rho I \succ 0_n$; see [15, 26] for a similar perturbation technique.

We have the following lemma, analogous to Lemma 2.1, showing that d is a descent direction at x whenever $d \neq 0$. We include its proof for completeness.

Lemma 3.1 *For any $x \in X$, nonempty $\mathcal{J} \subseteq \mathcal{N}$ and symmetric $H \in \mathbb{R}^{n\times n}$ with $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succ 0_n$, let $d = d_H(x;\mathcal{J})$ and $g = \nabla f(x)$. Then*

$$g^T d \le -d^T H d \le -\lambda_{\min}(B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}})\|d\|^2. \quad (3.2)$$

Proof. For any $\alpha \in (0,1)$, we have from (3.1) and the convexity of the set X that

$$g^T d + \frac{1}{2}d^T H d \;\le\; g^T(\alpha d) + \frac{1}{2}(\alpha d)^T H(\alpha d) \;=\; \alpha g^T d + \frac{1}{2}\alpha^2 d^T H d.$$

Rearranging terms yields

$$(1-\alpha)g^T d + \frac{1}{2}(1-\alpha^2)d^T H d \le 0.$$

Since $1-\alpha^2 = (1-\alpha)(1+\alpha)$, dividing both sides by $1-\alpha > 0$ and then taking $\alpha \uparrow 1$ prove the first inequality in (3.2). Since $d_{\mathcal{J}} \in$ Null$(A_{\mathcal{J}})$ so that $d_{\mathcal{J}} = B_{\mathcal{J}} y$ for some vector y, we have

$$d^T H d = y^T B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} y \ge \|y\|^2 \lambda_{\min}(B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}) = \|d\|^2 \lambda_{\min}(B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}),$$

where the second equality uses $B_{\mathcal{J}}^T B_{\mathcal{J}} = I$. This proves the second inequality in (3.2).

■

We next choose a stepsize $\alpha > 0$ so that $x' = x + \alpha d$ achieves sufficient descent, and re-iterate. We now describe formally the block-coordinate gradient descent method.

CGD method:

Choose $x^0 \in X$. For $k = 0, 1, 2, ...$, generate x^{k+1} from x^k according to the iteration:

1. Choose a nonempty $\mathcal{J}^k \subseteq \mathcal{N}$ and a symmetric $H^k \in \mathbb{R}^{n \times n}$ with $B_{\mathcal{J}^k}^T H_{\mathcal{J}^k \mathcal{J}^k}^k B_{\mathcal{J}^k} \succ 0_n$.

2. Solve (3.1) with $x = x^k,\ \mathcal{J} = \mathcal{J}^k,\ H = H^k$ to obtain $d^k = d_{H^k}(x^k; \mathcal{J}^k)$.

3. Choose a stepsize $\alpha^k > 0$ and set $x^{k+1} = x^k + \alpha^k d^k$.

Various stepsize rules for smooth optimization [6, 32, 77] can be used in our setting. The following adaptation of the Armijo rule [6, page 225], based on Lemma 3.1 and Section 2.1, is simple and seems effective from both theoretical and practical standpoints.

Armijo rule:

Choose $\alpha_{\text{init}}^k > 0$ and let α^k be the largest element of $\{\alpha_{\text{init}}^k \beta^j\}_{j=0,1,...}$ satisfying

$$f(x^k + \alpha^k d^k) \leq f(x^k) + \sigma \alpha^k \Delta^k \quad \text{and} \quad x^k + \alpha^k d^k \in X, \tag{3.3}$$

where $0 < \beta < 1$, $0 < \sigma < 1$, $0 \leq \theta < 1$, and

$$\Delta^k \stackrel{\text{def}}{=} \nabla f(x^k)^T d^k + \theta d^{kT} H^k d^k. \tag{3.4}$$

Since $B_{\mathcal{J}^k}^T H_{\mathcal{J}^k \mathcal{J}^k}^k B_{\mathcal{J}^k} \succ 0_n$ and $0 \leq \theta < 1$, we see from Lemma 3.1 that

$$f(x^k + \alpha d^k) = f(x^k) + \alpha \nabla f(x^k)^T d^k + o(\alpha) \leq f(x^k) + \alpha \Delta^k + o(\alpha) \quad \forall \alpha \in (0, 1],$$

and $\Delta^k \leq (\theta - 1) d^{kT} H^k d^k < 0$ whenever $d^k \neq 0$. Since $0 < \sigma < 1$, this shows

that α^k given by the Armijo rule is well defined and positive. This rule, like that for sequential quadratic programming methods [6, 32, 77], requires only function evaluations. And, by choosing α^k_{init} based on the previous stepsize α^{k-1}, the number of function evaluations can be kept small in practice. Notice that Δ^k increases with θ. Thus, larger stepsizes will be accepted if we choose either σ near 0 or θ near 1. The minimization rule or the limited minimization rule [6, Section 2.2.1] (also see (3.20), (3.21)) can be used instead of the Armijo rule if the minimization is relatively inexpensive, such as for a QP.

For theoretical and practical efficiency, the working set $\mathcal{J}^k$ must be chosen judiciously so to ensure global convergence while balancing between convergence speed and the computational cost per iteration. Let us denote the optimal value of the direction subproblem (3.1) by

$$q_H(x;\mathcal{J}) \stackrel{\text{def}}{=} \left\{\nabla f(x)^T d + \frac{1}{2} d^T H d\right\}_{d=d_H(x;\mathcal{J})}. \tag{3.5}$$

Intuitively, $q_H(x;\mathcal{J})$ is the predicted descent when x is moved along the direction $d_H(x;\mathcal{J})$. We will choose the working set $\mathcal{J}^k$ to satisfy

$$q_{D^k}(x^k;\mathcal{J}^k) \leq \upsilon\ q_{D^k}(x^k;\mathcal{N}), \tag{3.6}$$

where $D^k \succ 0$ (typically diagonal) and $0 < \upsilon \leq 1$. (In fact, it suffices that $B_{\mathcal{N}}^T D^k B_{\mathcal{N}} \succ 0_n$ for our analysis.) This working set choice is motivated by the *Gauss-Southwell-q* rule in Chapter 2, which has good convergence properties in theory and in practice. It is similar in spirit to (1.11) with $\phi(\alpha) = \upsilon\alpha$, which corresponds to (3.6) with $m = 1$, $|\mathcal{J}^k| = 2$, $D^k = 0$, and X in (3.1) replaced by its tangent cone at x. We will discuss in Section 3.5 how to efficiently find a "small" working set $\mathcal{J}^k$ that satisfies (3.6) for some υ.

For the SVM QP (1.7), one choice of $\mathcal{J}^k$ that satisfies (3.6) with $\upsilon = 1/(n-\ell+1)$

is

$$\mathcal{J}^k \in \arg\min_{\mathcal{J}':|\mathcal{J}'|\leq\ell} \left\{ \begin{array}{ll} \min_d & \nabla f(x^k)^T d + \frac{1}{2} d^T \mathrm{diag}(Q) d \\ \text{s.t.} & a^T d = 0, \\ & 0 \leq x_j^k + d_j \leq C, \ j \in \mathcal{J}', \\ & d_j = 0, \ j \notin \mathcal{J}', \end{array} \right\} \tag{3.7}$$

where $\ell \in \{\mathrm{rank}(A)+1, \ldots, n\}$; see Proposition 3.2. However, no fast way to find such $\mathcal{J}^k$ is known.

3.2 Technical Preliminaries

In this section we study properties of the search direction $d_H(x;\mathcal{J})$ and the corresponding predicted descent $q_H(x;\mathcal{J})$. These will be useful for analyzing the global convergence and asymptotic convergence rate of the CGD method.

We say that an $x \in X$ is a *a stationary point* of f over X if $\nabla f(x)^T(y-x) \geq 0$ for all $y \in X$. This is equivalent to $d_D(x;\mathcal{N}) = 0$ for any $D \succ 0$; see [6, pages 229, 230].

The next lemma shows that $\|d_H(x;\mathcal{J})\|$ changes not too fast with the quadratic coefficients H. It will be used to prove Theorems 3.1 and 3.2. Recall that $B_\mathcal{J}$ is a matrix whose columns form an orthonormal basis for $\mathrm{Null}(A_\mathcal{J})$.

Lemma 3.2 *Fix any $x \in X$, nonempty $\mathcal{J} \subseteq \mathcal{N}$, and symmetric matrices $H, \tilde{H} \in \mathbb{R}^{n\times n}$ satisfying $U \succ 0_n$ and $\tilde{U} \succ 0_n$, where $U = B_\mathcal{J}^T H_{\mathcal{J}\mathcal{J}} B_\mathcal{J}$ and $\tilde{U} = B_\mathcal{J}^T \tilde{H}_{\mathcal{J}\mathcal{J}} B_\mathcal{J}$. Let $d = d_H(x;\mathcal{J})$ and $\tilde{d} = d_{\tilde{H}}(x;\mathcal{J})$. Then*

$$\|\tilde{d}\| \leq \frac{1 + \lambda_{\max}(S) + \sqrt{1 - 2\lambda_{\min}(S) + \lambda_{\max}(S)^2}}{2} \, \frac{\lambda_{\max}(U)}{\lambda_{\min}(\tilde{U})} \, \|d\|, \tag{3.8}$$

where $S = U^{-1/2}\tilde{U}U^{-1/2}$.

Proof. Since $d_j = \tilde{d}_j = 0$ for all $j \notin \mathcal{J}$, it suffices to prove the lemma for the case of $\mathcal{J} = \mathcal{N}$. Let $g = \nabla f(x)$. By the definition of d and $\tilde{d}$ and [90, Theorem 8.15],

$$d \in \arg\min_u \{(g + Hd)^T u \mid x + u \in X\},$$
$$\tilde{d} \in \arg\min_u \{(g + \tilde{H}\tilde{d})^T u \mid x + u \in X\}.$$

Thus

$$\begin{aligned}(g+Hd)^T d &\leq (g+Hd)^T \tilde{d},\\ (g+\tilde{H}\tilde{d})^T \tilde{d} &\leq (g+\tilde{H}\tilde{d})^T d.\end{aligned}$$

Adding the above two inequalities and rearranging terms yield

$$d^T H d - d^T(H+\tilde{H})\tilde{d} + \tilde{d}^T \tilde{H}\tilde{d} \leq 0.$$

Since $d, \tilde{d} \in \mathrm{Null}(A)$, we have $d = B_{\mathcal{N}} y$ and $\tilde{d} = B_{\mathcal{N}}\tilde{y}$ for some vectors $y, \tilde{y}$. Substituting these into the above inequality and using the definitions of $U, \tilde{U}$ yield

$$y^T U y - y^T(U+\tilde{U})\tilde{y} + \tilde{y}^T \tilde{U}\tilde{y} \leq 0.$$

Then proceeding as in the proof of Lemma 2.3 and using $\|d\| = \|y\|$, $\|\tilde{d}\| = \|\tilde{y}\|$ (since $B_{\mathcal{N}}^T B_{\mathcal{N}} = I$), we obtain (3.8). ∎

The next lemma gives a sufficient condition for the stepsize to satisfy the Armijo descent condition (3.3). This lemma will be used to prove Theorem 3.1(d). Its proof is similar to that of Lemma 2.5(b) and is included for completeness.

Lemma 3.3 *Suppose f satisfies*

$$\|\nabla f(y) - \nabla f(z)\| \leq L\|y - z\| \quad \forall y,\, z \in X, \tag{3.9}$$

for some $L \geq 0$. Fix any $x \in X$, nonempty $\mathcal{J} \subseteq \mathcal{N}$, and symmetric matrix $H \in \mathbb{R}^{n\times n}$ satisfying $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succeq \underline{\lambda} I$ with $\underline{\lambda} > 0$. Then, for any $\sigma \in (0,1)$, $\theta \in [0,1)$, and $0 \leq \alpha \leq 2\underline{\lambda}(1-\sigma+\sigma\theta)/L$ with $x + \alpha d \in X$, we have

$$f(x+\alpha d) - f(x) \leq \sigma\alpha(g^T d + \theta d^T H d), \tag{3.10}$$

where $d = d_H(x;\mathcal{J})$ and $g = \nabla f(x)$.

Proof. For any $\alpha \geq 0$ with $x+\alpha d \in X$, we have from the Cauchy-Schwarz inequality that

$$\begin{aligned} f(x+\alpha d) - f(x) &= \alpha g^T d + \int_0^1 (\nabla f(x+t\alpha d) - \nabla f(x))^T (\alpha d)\, dt \\ &\leq \alpha g^T d + \alpha \int_0^1 \|\nabla f(x+t\alpha d) - \nabla f(x)\| \|d\|\, dt \\ &\leq \alpha g^T d + \alpha^2 \frac{L}{2} \|d\|^2 \\ &= \alpha(g^T d + \theta d^T H d) - \alpha\theta d^T H d + \alpha^2 \frac{L}{2}\|d\|^2, \end{aligned} \tag{3.11}$$

where the third step uses (3.9) and $x+t\alpha d \in X$ when $0 \leq t \leq 1$. Since $\underline{\lambda}\|d\|^2 \leq d^T H d$ by Lemma 3.1, if in addition $\alpha \leq 2\underline{\lambda}(1-\sigma+\sigma\theta)/L$, then

$$\begin{aligned} \alpha\frac{L}{2}\|d\|^2 - \theta d^T H d &\leq (1-\sigma+\sigma\theta) d^T H d - \theta d^T H d \\ &= (1-\sigma)(1-\theta) d^T H d \\ &\leq -(1-\sigma)(\nabla f(x)^T d + \theta d^T H d), \end{aligned}$$

where the third step uses (3.2) in Lemma 3.1. This together with (3.11) yields (3.10). ■

The next lemma shows that $\nabla f(x)^T(x'-\bar{x})$ is bounded above by a weighted sum of $\|x-\bar{x}\|^2$ and $-q_D(x;\mathcal{J})$, where $x' = x+\alpha d$, $d = d_H(x;\mathcal{J})$, and $\mathcal{J}$ satisfies a condition analogous to (3.6). This lemma, which is new, will be needed to prove Theorem 3.2.

Lemma 3.4 *Fix any $x \in X$, nonempty $\mathcal{J} \subseteq \mathcal{N}$, and symmetric matrices $H, D \in \mathbb{R}^{n\times n}$ satisfying $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succ 0_n$, $\bar{\delta} I \succeq D \succ 0_n$, and*

$$q_D(x;\mathcal{J}) \leq \upsilon\ q_D(x;\mathcal{N}), \tag{3.12}$$

with $\bar{\delta} > 0$, $0 < \upsilon \leq 1$. Then, for any $\bar{x} \in X$ and $\alpha \geq 0$, we have

$$g^T(x'-\bar{x}) \leq \frac{\bar{\delta}}{2}\|\bar{x}-x\|^2 - \frac{1}{\upsilon} q_D(x;\mathcal{J}), \tag{3.13}$$

where $d = d_H(x;\mathcal{J})$, $g = \nabla f(x)$, and $x' = x+\alpha d$.

Proof. Since $\bar{x} - x$ is a feasible solution of the minimization subproblem (3.1) corresponding to $\mathcal{N}$ and D, we have

$$q_D(x;\mathcal{N}) \leq g^T(\bar{x} - x) + \frac{1}{2}(\bar{x} - x)^T D(\bar{x} - x).$$

Since $\bar{\delta} I \succeq D \succ 0_n$, we have $0 \leq (\bar{x} - x)^T D(\bar{x} - x) \leq \bar{\delta}\|\bar{x} - x\|^2$. This together with (3.12) yields

$$\frac{1}{\upsilon} q_D(x;\mathcal{J}) \leq g^T(\bar{x} - x) + \frac{\bar{\delta}}{2}\|\bar{x} - x\|^2.$$

Rearranging terms, we have

$$g^T(x - \bar{x}) \leq \frac{\bar{\delta}}{2}\|\bar{x} - x\|^2 - \frac{1}{\upsilon} q_D(x;\mathcal{J}). \tag{3.14}$$

By the definition of d and Lemma 3.1, we have $g^T d \leq 0$. Since $\alpha \geq 0$, this implies $\alpha g^T d \leq 0$. Adding this to (3.14) yields (3.13). ■

3.3 Global Convergence Analysis

In this section we analyze the global convergence of the CGD method under the following reasonable assumption on our choice of H^k.

Assumption 3.1 *$\bar{\lambda} I \succeq B^T_{\mathcal{J}^k} H^k_{\mathcal{J}^k \mathcal{J}^k} B_{\mathcal{J}^k} \succeq \underline{\lambda} I$ for all k, where $0 < \underline{\lambda} \leq \bar{\lambda}$.*

First, we have the following lemma relating the optimal solution and the optimal objective value of (3.1) when $\mathcal{J} = \mathcal{J}^k$ and $H = D^k$. This lemma will be used to prove Theorem 3.1(c).

Lemma 3.5 *For any $x^k \in X$, nonempty $\mathcal{J}^k \subseteq \mathcal{N}$, and $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ ($0 < \underline{\delta} \leq \bar{\delta}$), $k = 0, 1, \ldots$, if $\{x^k\}$ is convergent, then $\{d_{D^k}(x^k;\mathcal{J}^k)\} \to 0$ if and only if $\{q_{D^k}(x^k;\mathcal{J}^k)\} \to 0$.*

Proof. Let $\{x^k\}$ be a convergent sequence in X. Then $\{\nabla f(x^k)\}$ is convergent by the continuity of ∇f. If $\{d_{D^k}(x^k;\mathcal{J}^k)\} \to 0$, then (3.5) and the boundedness of $\{D^k\}$

imply $\{q_{D^k}(x^k;\mathcal{J}^k)\} \to 0$. Conversely, we have from (3.5) and (3.2) with $H = D^k$ that $q_{D^k}(x^k;\mathcal{J}^k) \le -\frac{1}{2} d_{D^k}(x^k;\mathcal{J}^k)^T D^k d_{D^k}(x^k;\mathcal{J}^k) \le -\frac{\delta}{2}\|d_{D^k}(x^k;\mathcal{J}^k)\|^2$ for all k. Thus if $\{q_{D^k}(x^k;\mathcal{J}^k)\} \to 0$, then $\{d_{D^k}(x^k;\mathcal{J}^k)\} \to 0$. ■

Using Lemmas 3.1, 3.2, 3.3, and 3.5, we have the following global convergence result, under Assumption 3.1, for the CGD method with $\{\mathcal{J}^k\}$ chosen by the Gauss-Southwell rule (3.6) and $\{\alpha^k\}$ chosen by the Armijo rule (3.3). Its proof adapts the analysis of gradient methods for unconstrained smooth optimization [6, pages 43-45] to handle constraints and block-coordinate updating.

Theorem 3.1 *Let $\{x^k\}$, $\{\mathcal{J}^k\}$, $\{H^k\}$, $\{d^k\}$ be sequences generated by the CGD method under Assumption 3.1, where $\{\alpha^k\}$ is chosen by the Armijo rule with $\inf_k \alpha^k_{\text{init}} > 0$. Then the following results hold.*

(a) *$\{f(x^k)\}$ is nonincreasing and Δ^k given by (3.4) satisfies*

$$-\Delta^k \ge (1-\theta) d^{k^T} H^k d^k \ge (1-\theta)\underline{\lambda}\|d^k\|^2 \quad \forall k, \tag{3.15}$$

$$f(x^{k+1}) - f(x^k) \le \sigma\alpha^k\Delta^k \le 0 \quad \forall k. \tag{3.16}$$

(b) *If $\{x^k\}_{\mathcal{K}}$ is a convergent subsequence of $\{x^k\}$, then $\{\alpha^k\Delta^k\} \to 0$ and $\{d^k\}_{\mathcal{K}} \to 0$. If in addition $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k, where $0 < \underline{\delta} \le \bar{\delta}$, then $\{d_{D^k}(x^k;\mathcal{J}^k)\}_{\mathcal{K}} \to 0$.*

(c) *If $\{\mathcal{J}^k\}$ is chosen by (3.6) and $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k, where $0 < \underline{\delta} \le \bar{\delta}$, then every cluster point of $\{x^k\}$ is a stationary point of (1.6).*

(d) *If f satisfies (3.9) for some $L \ge 0$, then $\inf_k \alpha^k > 0$. If $\lim_{k\to\infty} f(x^k) > -\infty$ also, then $\{\Delta^k\} \to 0$ and $\{d^k\} \to 0$.*

Proof. (a) The first inequality in (3.15) follows from (3.4) and Lemma 3.1. The second inequality follows from $0 \le \theta < 1$, Lemma 3.1, and $\lambda_{\min}(B^T_{\mathcal{J}^k} H^k_{\mathcal{J}^k\mathcal{J}^k} B_{\mathcal{J}^k}) \ge \bar{\lambda}$.

Since $x^{k+1} = x^k + \alpha^k d^k$ and α^k is chosen by the Armijo rule (3.3), we have (3.16) and hence $\{f(x^k)\}$ is nonincreasing.

(b) Let $\{x^k\}_{\mathcal{K}}$ ($\mathcal{K} \subseteq \{0, 1, \ldots\}$) be a subsequence of $\{x^k\}$ converging to some $\bar{x}$. Since f is smooth, $f(\bar{x}) = \lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} f(x^k)$. Since $\{f(x^k)\}$ is nonincreasing, this implies that $\{f(x^k)\} \downarrow f(\bar{x})$. Hence, $\{f(x^k) - f(x^{k+1})\} \to 0$. Then, by (3.16),

$$\{\alpha^k \Delta^k\} \to 0. \tag{3.17}$$

Suppose that $\{d^k\}_{\mathcal{K}} \not\to 0$. By passing to a subsequence if necessary, we can assume that, for some $\delta > 0$, $\|d^k\| \geq \delta$ for all $k \in \mathcal{K}$. Then, by (3.15) and (3.17), $\{\alpha^k\}_{\mathcal{K}} \to 0$. Since $\inf_k \alpha^k_{\text{init}} > 0$, there exists some index $\bar{k} \geq 0$ such that $\alpha^k < \alpha^k_{\text{init}}$ and $\alpha^k \leq \beta$ for all $k \in \mathcal{K}$ with $k \geq \bar{k}$. Since $x^k + d^k \in X$ and X is convex, the latter implies that $x^k + (\alpha^k/\beta)d^k \in X$ for all $k \in \mathcal{K}$ with $k \geq \bar{k}$. Since α^k is chosen by the Armijo rule, this in turn implies that

$$f(x^k + (\alpha^k/\beta)d^k) - f(x^k) > \sigma(\alpha^k/\beta)\Delta^k \quad \forall k \in \mathcal{K},\ k \geq \bar{k}.$$

Using the definition of Δ^k, we can rewrite this as

$$-(1-\sigma)\Delta^k + \theta {d^k}^T H^k d^k < \frac{f(x^k + (\alpha^k/\beta)d^k) - f(x^k)}{\alpha^k/\beta} - \nabla f(x^k)^T d^k \quad \forall k \in \mathcal{K},\ k \geq \bar{k}.$$

By (3.15), the left-hand side is greater than or equal to $((1-\sigma)(1-\theta)+\theta)\underline{\lambda}\|d^k\|^2$, so dividing both sides by $\|d^k\|$ yields

$$((1-\sigma)(1-\theta)+\theta)\underline{\lambda}\|d^k\| < \frac{f(x^k + \hat{\alpha}^k d^k/\|d^k\|) - f(x^k)}{\hat{\alpha}^k} - \frac{\nabla f(x^k)^T d^k}{\|d^k\|} \quad \forall k \in \mathcal{K},\ k \geq \bar{k}, \tag{3.18}$$

where we let $\hat{\alpha}^k = \alpha^k\|d^k\|/\beta$. By (3.15), $-\alpha^k\Delta^k \geq (1-\theta)\underline{\lambda}\alpha^k\|d^k\|^2 \geq (1-\theta)\underline{\lambda}\alpha^k\|d^k\|\delta$ for all $k \in \mathcal{K}$, so (3.17) and $(1-\theta)\underline{\lambda} > 0$ imply $\{\alpha^k\|d^k\|\}_{\mathcal{K}} \to 0$ and hence $\{\hat{\alpha}^k\}_{\mathcal{K}} \to 0$. Also, since $\{d^k/\|d^k\|\}_{\mathcal{K}}$ is bounded, by passing to a subsequence if necessary, we can assume that $\{d^k/\|d^k\|\}_{\mathcal{K}} \to$ some $\bar{d}$. Taking the limit as $k \in \mathcal{K}, k \to \infty$ in the inequality (3.18) and using the smoothness of f, we obtain

$$0 < ((1-\sigma)(1-\theta)+\theta)\underline{\lambda}\delta \leq \nabla f(\bar{x})^T\bar{d} - \nabla f(\bar{x})^T\bar{d} = 0,$$

a clear contradiction. Thus $\{d^k\}_{\mathcal{K}} \to 0$.

Suppose that, in addition, $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k. Let $\tilde{U}^k = B_{\mathcal{J}^k}^T D_{\mathcal{J}^k \mathcal{J}^k}^k B_{\mathcal{J}^k}$ and $U^k = B_{\mathcal{J}^k}^T H_{\mathcal{J}^k \mathcal{J}^k}^k B_{\mathcal{J}^k}$. Then, for each k, $\bar{\delta} I \succeq \tilde{U}^k \succeq \underline{\delta} I$ (since $B_{\mathcal{J}^k}^T B_{\mathcal{J}^k} = I$) as well as $\bar{\lambda} I \succeq U^k \succeq \underline{\lambda} I$. Then

$$\frac{\bar{\delta}}{\underline{\lambda}} I \succeq \bar{\delta} (U^k)^{-1} \succeq (U^k)^{-1/2} \tilde{U}^k (U^k)^{-1/2} \succeq \underline{\delta} (U^k)^{-1} \succeq \frac{\underline{\delta}}{\bar{\lambda}} I,$$

so (3.8) in Lemma 3.2 yields

$$\|d_{D^k}(x^k; \mathcal{J}^k)\| \leq \frac{1 + \bar{\delta}/\underline{\lambda} + \sqrt{1 - 2\underline{\delta}/\bar{\lambda} + (\bar{\delta}/\underline{\lambda})^2}}{2} \, \frac{\bar{\lambda}}{\underline{\delta}} \, \|d^k\|. \tag{3.19}$$

Since $\{d^k\}_{\mathcal{K}} \to 0$, this implies $\{d_{D^k}(x^k; \mathcal{J}^k)\}_{\mathcal{K}} \to 0$.

(c) Suppose that $\{\mathcal{J}^k\}$ is chosen by (3.6) and $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k and $\bar{x}$ is a cluster point of $\{x^k\}$. Let $\{x^k\}_{\mathcal{K}}$ be a subsequence of $\{x^k\}$ converging to $\bar{x}$. By (b), $\{d^k\}_{\mathcal{K}} \to 0$ and $\{d_{D^k}(x^k; \mathcal{J}^k)\}_{\mathcal{K}} \to 0$. By Lemma 3.5, $\{q_{D^k}(x^k; \mathcal{J}^k)\}_{\mathcal{K}} \to 0$. Since $\mathcal{J}^k$ satisfies (3.6), this implies that $\{q_{D^k}(x^k; \mathcal{N})\}_{\mathcal{K}} \to 0$. This together with Lemma 3.5 yields $\{d_{D^k}(x^k; \mathcal{N})\}_{\mathcal{K}} \to 0$.

By Lemma 3.2 with $\mathcal{J} = \mathcal{N}$, $H = D^k$, and $\tilde{H} = I$, we have

$$\|d_I(x^k; \mathcal{N})\| \leq \frac{1 + 1/\underline{\delta} + \sqrt{1 - 2/\bar{\delta} + (1/\underline{\delta})^2}}{2} \, \bar{\delta} \, \|d_{D^k}(x^k; \mathcal{N})\| \quad \forall k.$$

Hence $\{d_I(x^k; \mathcal{N})\}_{\mathcal{K}} \to 0$. A continuity argument then yields that $d_I(\bar{x}; \mathcal{N}) = 0$, so $\bar{x}$ is a stationary point of (1.6).

(d) Since α^k is chosen by the Armijo rule, either $\alpha^k = \alpha_{\text{init}}^k$ or else, by Lemma 3.3 and $x^k + d^k \in X$, $\alpha^k/\beta > \min\{1, 2\underline{\lambda}(1 - \sigma + \sigma\theta)/L\}$. Since $\inf_k \alpha_{\text{init}}^k > 0$, this implies $\inf_k \alpha^k > 0$. If $\lim_{k \to \infty} f(x^k) > -\infty$ also, then this and (3.16) imply $\{\Delta^k\} \to 0$, which together with (3.15) imply $\{d^k\} \to 0$. ∎

Similar to the observation in [6, page 45], Theorem 3.1 readily extends to any stepsize rule that yields a larger descent than the Armijo rule at each iteration.

Corollary 3.1 *Theorem 3.1 still holds if in the CGD method the iterates are instead updated by* $x^{k+1} = x^k + \tilde{\alpha}^k d^k$, *where* $\tilde{\alpha}^k \geq 0$ *satisfies* $f(x^k + \tilde{\alpha}^k d^k) \leq f(x^k + \alpha^k d^k)$ *and* $x^k + \tilde{\alpha}^k d^k \in X$ *for* $k = 0, 1, \ldots$, *and* $\{\alpha^k\}$ *is chosen by the Armijo rule with* $\inf_k \alpha^k_{\text{init}} > 0$.

Proof. It is readily seen using $f(x^{k+1}) \leq f(x^k + \alpha^k d^k)$ that Theorem 3.1(a) holds. The proofs of Theorem 3.1(b)–(d) remain unchanged. ■

For example, $\tilde{\alpha}^k$ may be generated by the minimization rule:

$$\tilde{\alpha}^k \in \arg\min_{\alpha \geq 0} \{f(x^k + \alpha d^k) \mid x^k + \alpha d^k \in X\} \tag{3.20}$$

or by the limited minimization rule:

$$\tilde{\alpha}^k \in \arg\min_{0 \leq \alpha \leq s} \{f(x^k + \alpha d^k) \mid x^k + \alpha d^k \in X\}, \tag{3.21}$$

where $0 < s < \infty$. The latter stepsize rule yields a larger descent than the Armijo rule with $\alpha^k_{\text{init}} = s$. We will use the minimization rule in our numerical tests on SVM QP; see Section 3.6.

3.4 Convergence Rate Analysis

In this section we analyze the asymptotic convergence rate of the CGD method under the following reasonable assumption; see [63]. In what follows, $\bar{X}$ denotes the set of stationary points of (1.6) and

$$\text{dist}(x, \bar{X}) \stackrel{\text{def}}{=} \min_{\bar{x} \in \bar{X}} \|x - \bar{x}\| \qquad \forall x \in \mathbb{R}^n.$$

Assumption 3.2 (a) $\bar{X} \neq \emptyset$ *and, for any* $\zeta \geq \min_{x \in X} f(x)$, *there exist scalars* $\tau > 0$ *and* $\epsilon > 0$ *such that*

$$\text{dist}(x, \bar{X}) \leq \tau \|d_I(x; \mathcal{N})\| \quad \textit{whenever} \quad x \in X,\ f(x) \leq \zeta,\ \|d_I(x; \mathcal{N})\| \leq \epsilon.$$

(b) *There exists a scalar $\rho > 0$ such that*

$$\|x - y\| \geq \rho \quad \textit{whenever} \quad x \in \bar{X},\ y \in \bar{X},\ f(x) \neq f(y).$$

Assumption 3.2 is identical to Assumptions A and B in [63]. Assumption 3.2(b) says that the isocost surfaces of f restricted to the solution set $\bar{X}$ are "properly separated." Assumption 3.2(b) holds automatically if f is a convex function. It also holds if f is quadratic and X is polyhedral [60, Lemma 3.1]. Assumption 3.2(a) is a local Lipschitzian error bound assumption, saying that the distance from x to $\bar{X}$ is locally in the order of the norm of the residual at x. Error bounds of this kind have been extensively studied.

Since X is polyhedral, we immediately have from [63, Theorem 2.1] the following sufficient conditions for Assumption 3.2(a) to hold. In particular, Assumption 3.2(a) and (b) hold for (1.7) and, more generally, any QP [60, 63].

Proposition 3.1 *Suppose that $\bar{X} \neq \emptyset$ and any of the following conditions hold.*

C1 *f is strongly convex and ∇f is Lipschitz continuous on X (i.e., (3.9) holds for some $L \geq 0$).*

C2 *f is quadratic.*

C3 *$f(x) = g(Ex) + q^T x$ for all $x \in \mathbb{R}^n$, where $E \in \mathbb{R}^{m \times n}$, $q \in \mathbb{R}^n$, and g is a strongly convex differentiable function on $\mathbb{R}^m$ with ∇g Lipschitz continuous on $\mathbb{R}^m$.*

C4 *$f(x) = \max_{y \in Y}\{(Ex)^T y - g(y)\} + q^T x$ for all $x \in \mathbb{R}^n$, where Y is a polyhedral set in $\mathbb{R}^m$, $E \in \mathbb{R}^{m \times n}$, $q \in \mathbb{R}^n$, and g is a strongly convex differentiable function on $\mathbb{R}^m$ with ∇g Lipschitz continuous on $\mathbb{R}^m$.*

Then Assumption 3.2(a) holds.

Using Theorem 3.1 and Lemmas 3.1, 3.2, and 3.4, we have the following linear convergence result, under Assumptions 3.1, 3.2, and (3.9), for the CGD method with $\{\mathcal{J}^k\}$ chosen by (3.6) and $\{\alpha^k\}$ chosen by the Armijo rule. Its proof adapts that of Theorem 2.4 to constrained problems. To our knowledge, this is the first linear convergence result for a block-coordinate update method for general linearly constrained smooth optimization. Moreover, it does not assume f is strongly convex or the stationary points satisfy strict complementarity.

Theorem 3.2 *Assume that f satisfies* (3.9) *for some $L \geq 0$ and Assumption 3.2. Let $\{x^k\}$, $\{H^k\}$, $\{d^k\}$ be sequences generated by the CGD method satisfying Assumption 3.1, where $\{\mathcal{J}^k\}$ is chosen by* (3.6), *$\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k ($0 < \underline{\delta} \leq \bar{\delta}$), and $\{\alpha^k\}$ is chosen by the Armijo rule with $\sup_k \alpha^k_{\rm init} < \infty$ and $\inf_k \alpha^k_{\rm init} > 0$. Then either $\{f(x^k)\} \downarrow -\infty$ or $\{f(x^k)\}$ converges at least Q-linearly and $\{x^k\}$ converges at least R-linearly to a point in $\bar{X}$.*

Proof. For each $k = 0, 1, \ldots$, (3.4) and $d^k = d_{H^k}(x^k; \mathcal{J}^k)$ imply that

$$\begin{aligned} \Delta^k + \left(\frac{1}{2} - \theta\right) d^{k^T} H^k d^k &= g^{k^T} d^k + \frac{1}{2} d^{k^T} H^k d^k \\ &\leq g^{k^T} \tilde{d}^k + \frac{1}{2} (\tilde{d}^k)^T H^k \tilde{d}^k \\ &= q_{D^k}(x^k; \mathcal{J}^k) + \frac{1}{2} (\tilde{d}^k)^T (H^k - D^k) \tilde{d}^k \\ &\leq q_{D^k}(x^k; \mathcal{J}^k) + \omega \|d^k\|^2, \end{aligned} \tag{3.22}$$

where we let $g^k = \nabla f(x^k)$ and $\tilde{d}^k = d_{D^k}(x^k; \mathcal{J}^k)$, and the last step uses (3.19) and $(\tilde{d}^k)^T (H^k - D^k) \tilde{d}^k \leq (\bar{\lambda} - \underline{\delta}) \|\tilde{d}^k\|^2$. Here, $\omega \in \mathbb{R}$ is a constant depending on $\bar{\lambda}, \underline{\lambda}, \bar{\delta}, \underline{\delta}$ only. Also, by (3.5) and Lemma 3.1 with $\mathcal{J} = \mathcal{N}$, $H = D^k$, we have

$$\begin{aligned} q_{D^k}(x^k; \mathcal{N}) &= \left(g^{k^T} d + \frac{1}{2} d^T D^k d \right)_{d = d_{D^k}(x^k; \mathcal{N})} \\ &\leq \left(-\frac{1}{2} d^T D^k d \right)_{d = d_{D^k}(x^k; \mathcal{N})} \\ &\leq -\frac{\delta}{2} \|d_{D^k}(x^k; \mathcal{N})\|^2 \quad \forall k, \end{aligned} \tag{3.23}$$

where the last inequality uses $D^k \succeq \underline{\delta} I$.

By Theorem 3.1(a), $\{f(x^k)\}$ is nonincreasing. Thus either $\{f(x^k)\} \downarrow -\infty$ or $\lim_{k\to\infty} f(x^k) > -\infty$. Suppose the latter. Since α^k is chosen by the Armijo rule with $\inf_k \alpha^k_{\text{init}} > 0$, Theorem 3.1(d) implies $\{\Delta^k\} \to 0$ and $\{d^k\} \to 0$. Since $\{H^k\}$ is bounded by Assumption 3.1, we obtain from (3.22) that $0 \le \lim_{k\to\infty} \inf q_{D^k}(x^k; \mathcal{J}^k)$. Then (3.6) and (3.23) yield $\{d_{D^k}(x^k; \mathcal{N})\} \to 0$.

By Lemma 3.2 with $\mathcal{J} = \mathcal{N}$, $H = D^k$ and $\tilde{H} = I$, we have

$$\|d_I(x^k; \mathcal{N})\| \le \frac{1 + 1/\underline{\delta} + \sqrt{1 - 2/\bar{\delta} + (1/\underline{\delta})^2}}{2} \bar{\delta} \, \|d_{D^k}(x^k; \mathcal{N})\| \quad \forall k. \tag{3.24}$$

Hence $\{d_I(x^k; \mathcal{N})\} \to 0$. Since $\{f(x^k)\}$ is nonincreasing, so that $f(x^k) \le f(x^0)$, as well as $x^k \in X$, for all k. Then, by Assumption 3.2(a), there exist $\bar{k}$ and $\tau > 0$ such that

$$\|x^k - \bar{x}^k\| \le \tau \|d_I(x^k; \mathcal{N})\| \quad \forall k \ge \bar{k}, \tag{3.25}$$

where $\bar{x}^k \in \bar{X}$ satisfies $\|x^k - \bar{x}^k\| = \text{dist}(x^k, \bar{X})$. Since $\{d_I(x^k; \mathcal{N})\} \to 0$, this implies $\{x^k - \bar{x}^k\} \to 0$. Since $\{x^{k+1} - x^k\} = \{\alpha^k d^k\} \to 0$, this and Assumption 3.2(b) imply that $\{\bar{x}^k\}$ eventually settles down at some isocost surface of f, i.e., there exist an index $\hat{k} \ge \bar{k}$ and a scalar $\bar{v}$ such that

$$f(\bar{x}^k) = \bar{v} \quad \forall k \ge \hat{k}. \tag{3.26}$$

Fix any index $k \ge \hat{k}$. Since $\bar{x}^k$ is a stationary point of f over X, we have

$$\nabla f(\bar{x}^k)^T (x^k - \bar{x}^k) \ge 0.$$

We also have from the Mean Value Theorem that

$$f(x^k) - f(\bar{x}^k) = \nabla f(\psi^k)^T (x^k - \bar{x}^k),$$

for some ψ^k lying on the line segment joining x^k with $\bar{x}^k$. Since x^k, $\bar{x}^k$ lie in the convex set X, so does ψ^k. Combining these two relations and using (3.26), we obtain

$$\bar{v} - f(x^k) \;\; \le \;\; (\nabla f(\bar{x}^k) - \nabla f(\psi^k))^T (x^k - \bar{x}^k)$$

$$\begin{aligned} &\leq \|\nabla f(\bar{x}^k) - \nabla f(\psi^k)\|\|x^k - \bar{x}^k\| \\ &\leq L\|x^k - \bar{x}^k\|^2, \end{aligned}$$

where the last inequality uses (3.9) and $\|\psi^k - \bar{x}^k\| \leq \|x^k - \bar{x}^k\|$. This together with $\{x^k - \bar{x}^k\} \to 0$ proves that

$$\liminf_{k\to\infty} f(x^k) \geq \bar{v}. \tag{3.27}$$

For each index $k \geq \hat{k}$, we have from (3.26) that

$$\begin{aligned} f(x^{k+1}) - \bar{v} &= f(x^{k+1}) - f(\bar{x}^k) \\ &= \nabla f(\tilde{x}^k)^T(x^{k+1} - \bar{x}^k) \\ &= (\nabla f(\tilde{x}^k) - g^k)^T(x^{k+1} - \bar{x}^k) + g^{k^T}(x^{k+1} - \bar{x}^k) \\ &\leq L\|\tilde{x}^k - x^k\|\|x^{k+1} - \bar{x}^k\| + \frac{\bar{\delta}}{2}\|x^k - \bar{x}^k\|^2 - \frac{1}{\upsilon}q_{D^k}(x^k; \mathcal{J}^k), \end{aligned} \tag{3.28}$$

where the second step uses the Mean Value Theorem with $\tilde{x}^k$ a point lying on the segment joining x^{k+1} with $\bar{x}^k$ (so that $\tilde{x}^k \in X$); the fourth step uses (3.9) and Lemma 3.4. Using the inequalities $\|\tilde{x}^k - x^k\| \leq \|x^{k+1} - x^k\| + \|x^k - \bar{x}^k\|$, $\|x^{k+1} - \bar{x}^k\| \leq \|x^{k+1} - x^k\| + \|x^k - \bar{x}^k\|$ and $\|x^{k+1} - x^k\| = \alpha^k\|d^k\|$, we see from (3.25), and $\sup_k \alpha^k < \infty$ (since $\sup_k \alpha^k_{\text{init}} < \infty$) that the right-hand side of (3.28) is bounded above by

$$C_1\left(\|d^k\|^2 - q_{D^k}(x^k; \mathcal{J}^k) + \|d_I(x^k; \mathcal{N})\|^2\right) \tag{3.29}$$

for all $k \geq \hat{k}$, where $C_1 > 0$ is some constant depending on $L, \tau, \bar{\delta}, \upsilon, \sup_k \alpha^k$ only.

By (3.15), we have

$$\underline{\lambda}\|d^k\|^2 \leq d^{k^T}H^k d^k \leq -\frac{1}{1-\theta}\Delta^k \quad \forall k. \tag{3.30}$$

By (3.23) and (3.24), we also have

$$\|d_I(x^k; \mathcal{N})\|^2 \leq \left(1 + 1/\underline{\delta} + \sqrt{1 - 2/\bar{\delta} + (1/\underline{\delta})^2}\right)^2 \frac{\bar{\delta}^2}{2\underline{\delta}}\left(-q_{D^k}(x^k; \mathcal{N})\right) \quad \forall k.$$

Thus, the quantity in (3.29) is bounded above by

$$C_2\left(-\Delta^k - q_{D^k}(x^k; \mathcal{J}^k) - q_{D^k}(x^k; \mathcal{N})\right) \tag{3.31}$$

for all $k \geq \hat{k}$, where $C_2 > 0$ is some constant depending on $L, \tau, \bar{\delta}, \underline{\delta}, \theta, \underline{\lambda}, \upsilon, \sup_k \alpha^k$ only.

Combining (3.22) with (3.30) yields

$$\begin{aligned} -q_{D^k}(x^k; \mathcal{J}^k) &\leq -\Delta^k + \left(\theta - \frac{1}{2}\right) d^{k^T} H^k d^k + \omega \|d^k\|^2 \\ &\leq -\Delta^k - \max\left\{0, \theta - \frac{1}{2}\right\} \frac{1}{1-\theta} \Delta^k - \frac{\omega}{\underline{\lambda}(1-\theta)} \Delta^k. \end{aligned} \tag{3.32}$$

Combining (3.6) and (3.32), we see that the quantity in (3.31) is bounded above by

$$-C_3 \Delta^k$$

all $k \geq \hat{k}$, where $C_3 > 0$ is some constant depending on $L, \tau, \bar{\delta}, \underline{\delta}, \theta, \bar{\lambda}, \underline{\lambda}, \upsilon, \sup_k \alpha^k$ only. Thus the right-hand side of (3.28) is bounded above by $-C_3 \Delta^k$ for all $k \geq \hat{k}$. Combining this with (3.16), (3.28), and $\inf_k \alpha^k > 0$ (see Theorem 3.1(d)) yields

$$f(x^{k+1}) - \bar{v} \leq C_4 (f(x^k) - f(x^{k+1})) \quad \forall k \geq \hat{k},$$

where $C_4 = C_3 / (\sigma \inf_k \alpha^k)$. Upon rearranging terms and using (3.27), we have

$$0 \leq f(x^{k+1}) - \bar{v} \leq \frac{C_4}{1 + C_4} (f(x^k) - \bar{v}) \quad \forall k \geq \hat{k},$$

so $\{f(x^k)\}$ converges to $\bar{v}$ at least Q-linearly.

Finally, by (3.16), (3.30), and $x^{k+1} - x^k = \alpha^k d^k$, we have

$$\sigma(1-\theta)\underline{\lambda} \frac{\|x^{k+1} - x^k\|^2}{\alpha^k} \leq f(x^k) - f(x^{k+1}) \quad \forall k \geq \hat{k}.$$

This implies

$$\|x^{k+1} - x^k\| \leq \sqrt{\frac{\sup_k \alpha^k}{\sigma(1-\theta)\underline{\lambda}} (f(x^k) - f(x^{k+1}))} \quad \forall k \geq \hat{k}.$$

Since $\{f(x^k) - f(x^{k+1})\} \to 0$ at least R-linearly and $\sup_k \alpha^k < \infty$, this implies that $\{x^k\}$ converges at least R-linearly. ∎

Similar to Corollary 3.1, Theorem 3.2 readily extends to any stepsize rule that yields a uniformly bounded stepsize and a larger descent than the Armijo rule at each iteration. An example is the limited minimization rule (3.21).

Corollary 3.2 *Theorem 3.2 still holds if in the CGD method the iterates are instead updated by* $x^{k+1} = x^k + \tilde{\alpha}^k d^k$, *where* $\tilde{\alpha}^k \geq 0$ *satisfies* $\sup_k \tilde{\alpha}^k < \infty$, $f(x^k + \tilde{\alpha}^k d^k) \leq f(x^k + \alpha^k d^k)$ *and* $x^k + \tilde{\alpha}^k d^k \in X$ *for* $k = 0, 1, \ldots$, *and* $\{\alpha^k\}$ *is chosen by the Armijo rule with* $\sup_k \alpha^k_{\text{init}} < \infty$ *and* $\inf_k \alpha^k_{\text{init}} > 0$.

Proof. The only change to the proof of Theorem 3.2 is in proving (3.29) and the last paragraph, where we use $\|x^{k+1} - x^k\| = \tilde{\alpha}^k \|d^k\|$ and $\sup_k \tilde{\alpha}^k < \infty$ instead. ∎

3.5 Working Set Selection

In the previous two sections, we showed that the CGD method with $\mathcal{J}^k$ satisfying (3.6) has desirable convergence properties. The iteration complexity of this method depends on $|\mathcal{J}^k|$ and the complexity of finding $\mathcal{J}^k$. In this section we show that a "small" $\mathcal{J}^k$ satisfying (3.6), for some constant $0 < \upsilon \leq 1$, can be found "reasonably fast" when D^k is diagonal. Our approach is based on the notion of a conformal realization [87], [89, Section 10B] of $d_{D^k}(x^k, \mathcal{N})$. Specifically, for any $d \in \mathbb{R}^n$, the support of d is $\text{supp}(d) \stackrel{\text{def}}{=} \{j \in \mathcal{N} \mid d_j \neq 0\}$. A $d' \in \mathbb{R}^n$ is *conformal* to $d \in \mathbb{R}^n$ if

$$\text{supp}(d') \subseteq \text{supp}(d), \qquad d'_j d_j \geq 0 \; \forall j \in \mathcal{N}, \tag{3.33}$$

i.e., the nonzero components of d' have the same signs as the corresponding components of d. A nonzero $d \in \mathbb{R}^n$ is an *elementary vector* of $\text{Null}(A)$ if $d \in \text{Null}(A)$ and there is no nonzero $d' \in \text{Null}(A)$ that is conformal to d and $\text{supp}(d') \neq \text{supp}(d)$. Each elementary vector d satisfies $|\text{supp}(d)| \leq \text{rank}(A) + 1$ (since any subset of $\text{rank}(A) + 1$ columns of A are linearly dependent) [89, Exercise 10.6].

Proposition 3.2 *For any* $x \in X$, $\ell \in \{\text{rank}(A) + 1, \ldots, n\}$, *and diagonal* $D \succ 0$, *there exists a nonempty* $\mathcal{J} \subseteq \mathcal{N}$ *satisfying* $|\mathcal{J}| \leq \ell$ *and*

$$q_D(x; \mathcal{J}) \leq \frac{1}{n - \ell + 1} q_D(x; \mathcal{N}). \tag{3.34}$$

Proof. Let $d = d_D(x;\mathcal{N})$. We divide our argument into three cases.

Case (i) $d = 0$: Then $q_D(x;\mathcal{N}) = 0$. Thus, for any nonempty $\mathcal{J} \subseteq \mathcal{N}$ with $|\mathcal{J}| \leq \ell$, we have from (3.5) and Lemma 3.1 with $H = D$ that $q_D(x;\mathcal{J}) \leq 0 = q_D(x;\mathcal{N})$, so (3.34) holds.

Case (ii) $d \neq 0$ and $|\mathrm{supp}(d)| \leq \ell$: Then $\mathcal{J} = \mathrm{supp}(d)$ satisfies $q_D(x;\mathcal{J}) = q_D(x;\mathcal{N})$ and hence (3.34), as well as $|\mathcal{J}| \leq \ell$.

Case (iii) $d \neq 0$ and $|\mathrm{supp}(d)| > \ell$: Since $d \in \mathrm{Null}(A)$, it has a conformal realization [87], [89, Section 10B], namely,

$$d = v^1 + \cdots + v^s,$$

for some $s \geq 1$ and some nonzero elementary vectors $v^t \in \mathrm{Null}(A)$, $t = 1, \ldots, s$, conformal to d. Then for some $\alpha > 0$, $\mathrm{supp}(d')$ is a proper subset of $\mathrm{supp}(d)$ and $d' \in \mathrm{Null}(A)$, where $d' = d - \alpha v^1$. (Note that αv^1 is an elementary vector of $\mathrm{Null}(A)$, so that $|\mathrm{supp}(\alpha v^1)| \leq \mathrm{rank}(A) + 1 \leq \ell$.) We repeat the above reduction step with d' in place of d. Since $|\mathrm{supp}(d')| \leq |\mathrm{supp}(d)| - 1$, after at most $|\mathrm{supp}(d)| - \ell$ reduction steps, we obtain

$$d = d^1 + \cdots + d^r, \tag{3.35}$$

for some $r \leq |\mathrm{supp}(d)| - \ell + 1$ and some nonzero $d^t \in \mathrm{Null}(A)$ conformal to d with $|\mathrm{supp}(d^t)| \leq \ell$, $t = 1, \ldots, r$. Since $|\mathrm{supp}(d)| \leq n$, we have $r \leq n - \ell + 1$.

Since $l - x \leq d \leq u - x$, (3.35) and d^t being conformal to d imply $l - x \leq d^t \leq u - x$, $t = 1, \ldots, r$. Since $Ad^t = 0$, this implies $x + d^t \in X$, $t = 1, \ldots, r$. Also, (3.5) and (3.35) imply that

$$\begin{aligned} q_D(x;\mathcal{N}) &= g^T d + \frac{1}{2} d^T D d \\ &= \sum_{t=1}^{r} g^T d^t + \frac{1}{2} \sum_{s=1}^{r} \sum_{t=1}^{r} (d^s)^T D d^t \\ &\geq \sum_{t=1}^{r} g^T d^t + \frac{1}{2} \sum_{t=1}^{r} (d^t)^T D d^t \\ &\geq r \min_{t=1,\ldots,r} \left\{ g^T d^t + \frac{1}{2} (d^t)^T D d^t \right\}, \end{aligned}$$

where $g = \nabla f(x)$ and the first inequality uses (3.33) and $D \succ 0_n$ being diagonal, so that $(d^s)^T D d^t \geq 0$ for all s, t. Thus, if we let $\bar{t}$ be an index t attaining the above minimum and let $\mathcal{J} = \mathrm{supp}(d^{\bar{t}})$, then $|\mathcal{J}| \leq \ell$ and

$$\frac{1}{r} q_D(x; \mathcal{N}) \geq g^T d^{\bar{t}} + \frac{1}{2} (d^{\bar{t}})^T D d^{\bar{t}} \geq q_D(x; \mathcal{J}),$$

where the second inequality uses $x + d^{\bar{t}} \in X$ and $d^{\bar{t}}_j = 0$ for $j \notin \mathcal{J}$. ■

It can be seen from its proof that Proposition 3.2 still holds if the diagonal matrix D is only positive semidefinite, provided that $q_D(x; \mathcal{N}) > -\infty$ (such as when X is bounded). Thus Proposition 3.2 may be viewed as an extension of [11, Lemma 2.3] and [58, Theorem 2, part 2] for the case of $D = 0$.

The proof of Proposition 3.2 suggests, for any $\ell \in \{\mathrm{rank}(A)+1, \ldots, n\}$, an $O(n-\ell)$-step reduction procedure for finding a conformal realization (3.35) of $d = d_D(x; \mathcal{N})$ with $r \leq n - \ell + 1$ and a corresponding $\mathcal{J}$ satisfying $|\mathcal{J}| \leq \ell$ and (3.34).

- In the case of $m = 1$ and $\ell = 2$, by scaling A and dropping zero columns if necessary, we can without loss of generality assume that $A = e^T$ (so d has at least one positive and one negative component) and by recursively subtracting α from a positive component d_i and adding α to a negative component d_j, where $\alpha = \min\{d_i, -d_j\}$, we can find such a conformal realization in $O(n)$ operations.

- In the case of $m = 2$ and $\ell = 3$, the preceding procedure can be extended, by using sorting, to find such a conformal realization in $O(n \log n)$ operations. For brevity we omit the details.

- In general, each step of the reduction procedure requires finding a nonzero $v \in \mathrm{Null}(A)$ with $|\mathrm{supp}(v)| \leq \ell$ and conformal to a given $d \in \mathrm{Null}(A)$ with $|\mathrm{supp}(d)| > \ell$. This can be done in $O(m^3(n - \ell))$ operations as follows: Choose any $\mathcal{J} \subset \mathrm{supp}(d)$ with $|\mathcal{J}| = m + 1$. Find a nonzero $w \in \mathrm{Null}(A)$ with $w_j = 0$ for all $j \notin \mathcal{J}$. This can be done in $O(m^3)$ operations using Gaussian elimination.

Then for some $\alpha \in \mathbb{R}$, $\mathrm{supp}(d')$ is a proper subset of $\mathrm{supp}(d)$ and $d' \in \mathrm{Null}(A)$, where $d' = d - \alpha w$. Repeat this with d' in place of d. The number of repetitions is at most $\mathrm{supp}(d) - \ell \leq n - \ell$. The overall time complexity of this reduction procedure is $O(m^3(n-\ell)^2)$ operations.

For diagonal $D \succ 0$ and $m = 1$, $d_D(x; \mathcal{N})$ can be found by solving a continuous quadratic knapsack problem in $O(n)$ operations; see [8, 52] and references therein. For diagonal $D \succ 0$ and $m > 1$, $d_D(x; \mathcal{N})$ can be found using an algorithm described by Berman, Kovoor and Pardalos [4], which reportedly requires only $O(n)$ operations for each fixed m.

By combining the above observations, we conclude that, for $m = 1$ and $\ell = 2$, a working set $\mathcal{J}$ satisfying $|\mathcal{J}| \leq \ell$ and (3.34) can be found in $O(n)$ operations. For $m = 2$ and $\ell = 3$, such a working set $\mathcal{J}$ can be found in $O(n \log n)$ operations. For $m \geq 1$ and $\ell \in \{\mathrm{rank}(A) + 1, \ldots, n\}$, such a working set $\mathcal{J}$ can be found in $O(n^2)$ operations, where the constant in $O(\cdot)$ depends on m. It is an open question whether such a $\mathcal{J}$ can be found in $O(n)$ operations for a fixed $m \geq 2$.

3.6 Numerical Experience on SVM QP

In order to better understand its practical performance, we have implemented the CGD method in Fortran to solve the SVM QP (1.7)-(1.8), with the working set chosen as described in Section 3.5. In this case, the CGD method effectively reduces to an SMO method, so the novelty is our choice of the working set. In this section, we describe our implementation and report our numerical experience on some large two-class data classification problems. This is compared with LIBSVM (version 2.83), which chooses the working set differently, but with the same cardinality of 2.

In our tests, we use $C = 1, 10$ and the linear kernel $K(z_i, z_j) = z_i^T z_j$, the radial basis function kernel $K(z_i, z_j) = \exp(-\gamma \|z_i - z_j\|^2)$, the polynomial kernel $K(z_i, z_j) = (\gamma z_i^T z_j + s)^{deg}$, and the sigmoid kernel $K(z_i, z_j) = \tanh(\gamma z_i^T z_j + s)$ with $\gamma = 1/p$, $s = 0$,

Table 3.1: Comparing LIBSVM and CGD-3pair on large two-class data classification problems with linear kernel.

Data set	n/p	C	LIBSVM	CGD-3pair
			iter/obj/cpu	iter/obj/cpu(kcpu/gcpu)/kiter
a7a	16100/122	1	64108/-5699.253/1.3	56869/-5699.246/6.3(1.7/4.0)/21296
		10	713288/-56875.57/4.6	598827/-56875.55/59.4(20.3/34.1)/228004
a8a	22696/123	1	83019/-8062.410/2.7	95522/-8062.404/16.0(4.4/10.4)/35686
		10	663752/-80514.32/10.7	782559/-80514.27/106.2(35.1/61.2)/291766
a9a	32561/123	1	80980/-11433.38/5.7	110602/-11433.38/27.3(7.9/17.3)/40667
		10	1217122/-114237.4/24.0	1287193/-114237.4/291.4(92.9/175.8)/482716
ijcnn1	49990/22	1	16404/-8590.158/3.0	20297/-8590.155/6.5(2.2/4.0)/7870
		10	155333/-85441.01/4.2	155274/-85441.00/46.9(17.9/27.1)/63668
w7a	24692/300	1	66382/-765.4115/0.4	72444/-765.4116/8.2(2.5/5.4)/27920
		10	662877/-7008.306/1.1	626005/-7008.311/75.3(20.2/52.6)/241180

$deg = 3$ (cubic), the default setting for LIBSVM. For the sigmoid kernel, Q can be indefinite.

For the test problems, we use the two-class data classification problems from the LIBSVM data webpage (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/), for which $a \in \{-1, 1\}^n$. Due to memory limitation on our departmental Linux system, we limit n to at most 50,000 and p to at most 300. This yields the five problems shown in Tables 3.1 and 3.2.

Our implementation of the CGD method has the form

$$x^{k+1} = x^k + d_Q(x^k; \mathcal{J}^k), \quad k = 0, 1, \ldots,$$

with $|\mathcal{J}^k| = 2$ always. This corresponds to the CGD method with α^k chosen by the minimization rule. (The choice of H^k is actually immaterial here.) As with SMO methods, we initialize $x^0 = 0$ and, to save time, we cache the most recently used columns of Q, up to a user-specified limit `maxCN`, when updating the gradient $\nabla f(x^k) = Qx^k - e$. In our tests, we set `maxCN`=5000 for `ijcnn1` and otherwise `maxCN`=8000. We terminate the method when $-q_D(x^k; \mathcal{N}) \leq 10^{-5}$.

We describe below how we choose the working set $\mathcal{J}^k$ for the CGD method. We fix the diagonal scaling matrix

$$D = \operatorname{diag}\left[\max\{Q_{jj}, 10^{-5}\}\right]_{j=1,\ldots,n}.$$

Table 3.2: Comparing LIBSVM and CGD-3pair on large two-class data classification problems with nonlinear kernel.

Data set	n/p	C/kernel	**LIBSVM**	**CGD-3pair**
			iter/obj/cpu	iter/obj/cpu(kcpu/gcpu)/kiter
a7a	16100/122	1/rbf	4109/-5899.071/1.3	4481/-5899.070/1.0(0.1/0.8)/1593
		10/rbf	10385/-55195.29/1.4	16068/-55195.30/2.0(0.5/1.4)/5834
		1/poly	4149/-7720.475/1.1	4470/-7720.478/0.8(0.1/0.6)/1536
		10/poly	4153/-67778.17/1.2	4593/-67778.17/0.8(0.1/0.6)/1599
		1/sig	3941/-6095.529/1.7	4201/-6095.529/1.2(0.1/1.0)/1474
		10/sig	9942/-57878.56/1.7	10890/-57878.57/1.8(0.3/1.3)/4211
a8a	22696/123	1/rbf	5641/-8249.503/2.6	6293/-8249.504/2.1(0.2/1.6)/2222
		10/rbf	15469/-77831.16/2.7	26137/-77831.16/4.8(1.1/3.3)/9432
		1/poly	5819/-10797.56/2.2	6202/-10797.57/1.7(0.3/1.2)/2133
		10/poly	5656/-92870.58/2.1	6179/-92870.59/1.6(0.3/1.2)/2136
		1/sig	5473/-8491.386/3.2	6172/-8491.388/2.5(0.3/2.0)/2197
		10/sig	10955/-81632.40/3.3	17157/-81632.41/3.8(0.8/2.8)/6646
a9a	32561/123	1/rbf	7975/-11596.35/5.2	8863/-11596.35/4.3(0.5/3.3)/3110
		10/rbf	21843/-110168.5/5.4	36925/-110168.5/10.7(2.8/7.3)/13140
		1/poly	8282/-15243.50/4.5	8777/-15243.50/3.4(0.6/2.5)/3002
		10/poly	7816/-128316.3/4.0	8769/-128316.4/3.3(0.6/2.4)/3019
		1/sig	7363/-11904.90/6.5	8268/-11904.90/5.1(0.5/4.1)/2897
		10/sig	15944/-115585.1/6.4	15792/-115585.1/6.5(1.1/5.0)/5859
ijcnn1	49990/22	1/rbf	5713/-8148.187/4.6	6688/-8148.187/3.8(0.7/2.7)/2397
		10/rbf	6415/-61036.54/3.5	12180/-61036.54/4.8(1.3/3.2)/4570
		1/poly	5223/-9693.566/2.5	7156/-9693.620/3.1(0.9/2.0)/2580
		10/poly	5890/-95821.99/2.9	7987/-95822.02/3.3(1.0/2.1)/2949
		1/sig	6796/-9156.916/7.0	6856/-9156.916/5.0(0.8/3.9)/2452
		10/sig	10090/-88898.40/6.4	12420/-88898.39/6.5(1.4/4.7)/4975
w7a	24692/300	1/rbf	1550/-1372.011/0.4	1783/-1372.010/0.5(0.1/0.4)/731
		10/rbf	4139/-10422.69/0.4	4491/-10422.70/0.8(0.2/0.6)/1792
		1/poly	758/-1479.816/0.1	2297/-1479.825/0.5(0.1/0.4)/871
		10/poly	1064/-14782.40/0.2	3591/-14782.53/0.7(0.2/0.5)/1347
		1/sig	1477/-1427.453/0.4	2020/-1427.455/0.4(0.1/0.3)/796
		10/sig	2853/-11668.85/0.3	5520/-11668.86/0.9(0.2/0.6)/2205

(We also experimented with $D = I$, but this resulted in worse performance.) At the initial iteration and at certain subsequent iterations k, we compute $d_D(x^k, \mathcal{N})$ and $q_D(x^k; \mathcal{N})$ by using a linear-time Fortran code `k1vfo` provided to us by Krzysztof Kiwiel, as described in [52], to solve the corresponding continuous quadratic knapsack problem. Then we find a conformal realization of $d_D(x^k, \mathcal{N})$ using the linear-time reduction procedure described in Section 3.5. By Proposition 3.2, there exists at least one elementary vector in this realization whose support $\mathcal{J}$ satisfies

$$q_D(x^k; \mathcal{J}) \leq \frac{1}{n-1} q_D(x^k; \mathcal{N}).$$

From among all such $\mathcal{J}$, we find the best one (i.e., has the least $q_Q(x^k; \mathcal{J})$ value) and

make this the working set $\mathcal{J}^k$. (We also experimented with choosing one with the least $q_D(x^k; \mathcal{J})$ value, but this resulted in worse performance.) Since the continuous quadratic knapsack problem takes significant time to solve by `k1vfo`, we in addition find from among all such $\mathcal{J}$ the second-best and third-best ones, if they exist. (In our tests, they always exist.) If the second-best one is disjoint from $\mathcal{J}^k$, we make it the next working set $\mathcal{J}^{k+1}$, and if the third-best one is disjoint from both $\mathcal{J}^k$ and $\mathcal{J}^{k+1}$, we make it the second-next working set $\mathcal{J}^{k+2}$. (In our tests, the latter case occurs about 85-90% of the time.) If the second-best one is not disjoint from $\mathcal{J}^k$ but the third-best one is, then we make the third-best one the next working set $\mathcal{J}^{k+1}$. (We can also allow them to overlap, though the updating of $\nabla f(x^k)$ becomes more complicated and might not significantly improve the performance as the overlapping case occurs only about 10-15% of the time.) This working set selection procedure is then repeated at iteration $k+3$ or $k+2$ or $k+1$, depending on the case, and so on. It is straightforward to check that the global convergence and local linear convergence properties of the CGD method, as embodied in Theorems 3.1 and 3.2, extend to this choice of the working set. We refer to this CGD method as CGD-3pair.

We report in Tables 3.1 and 3.2 our numerical results, showing the number of iterations (iter), final f-value (obj), total time (cpu) in minutes. For CGD-3pair, we also show the total time taken by `k1vfo` to solve the knapsack problems (kcpu), the total time to compute/cache columns of Q and update the gradient (gcpu), and the total number of knapsack problems solved (kiter). All runs are performed on an HP DL360 workstation, running Red Hat Linux 3.5. LIBSVM and CGD-3pair are compiled using the Gnu C++ and F-77 compiler, respectively. From Tables 3.1 and 3.2, we see that the total number of iterations and the final f-value for CGD-3pair are comparable (within a factor of 2) to those of LIBSVM. On the other hand, the cpu times for CGD-3pair are much higher when the linear kernel is used, due to the greater times spent in `k1vfo` and for updating the gradient. When a nonlinear kernel is used, the cpu times for CGD-3pair are comparable to those of LIBSVM.

In general, CGD-3pair is significantly slower than LIBSVM when the linear kernel is used. But when a nonlinear kernel is used, CGD-3pair is comparable to LIBSVM in speed and solution quality. This suggests that the working set choice of Section 3.5 is a viable alternative to existing choices, especially when a nonlinear kernel is used. Conceivably CGD-3pair can be further speeded up by omitting infrequently updated components from computation ("shrinkage"), as is done in LIBSVM and SVM^{light}, and by incorporating "warm start" in the knapsack problem solver `k1vfo`, i.e., using a solution of the previous knapsack problem to initialize the solution of the next knapsack problem. Recoding CGD-3pair in C++ to make use of dynamic memory allocation and pointer structure is another direction for future research, as are extensions to multi-class data classification.

For the SVM QP (1.7), SMO method and CGD method have the advantage that they can be implemented to use only $O(n)$ operations per iteration and the number of iterations is typically $O(n)$ or lower. By starting at $x = 0$, the gradient can be computed in $O(n)$ operations and subsequently be updated in $O(n)$ operations. In contrast, an interior-point method would need to start at an $x > 0$, so it would take $O(n^2)$ operations just to compute the gradient, and then one needs to compute a quantity of the form $y^T(\rho I + Q)^{-1}y$ $(\rho > 0)$ at each iteration to obtain the search direction d. An exception is when Q has low rank r or is the sum of a rank-r matrix with a positive multiple of the identity matrix, such as linear SVM. Then Qx can be computed in $O(rn)$ operations and $(\rho I + Q)^{-1}y$ can be efficiently computed using low-rank updates [28, 29, 30].

3.7 Conclusions and Extensions

We have proposed a block-coordinate gradient descent method for linearly constrained smooth optimization, and have established its global convergence and asymptotic linear convergence to a stationary point under mild assumptions. On SVM QP (1.7), this method achieves linear convergence under no additional assumption, and is im-

plementable in $O(n)$ operations per iteration. Our preliminary numerical experience suggests that it can be competitive with state-of-the-art SVM code on large data classification problems when a nonlinear kernel is used.

There are many directions for future research. For example, in Section 3.5 we mentioned that a conformal realization can be found in $O(n \log n)$ operations when $m = 2$. However, for large-scale applications such as ν-SVM, this can still be slow. Can this be improved to $O(n)$ operations? Also, in our current implementation of the CGD method, we use a diagonal D^k when finding a working set $\mathcal{J}^k$ satisfying (3.6). Can we use a nondiagonal D^k and still efficiently find a $\mathcal{J}^k$ satisfying (3.6)?

The problem (1.1) and (1.6) can be generalized to the following problem:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) + cP(x) \\ \text{s.t.} \quad & Ax = b, \end{aligned}$$

where $c > 0$, $P : \mathbb{R}^n \to (-\infty, \infty]$ is a proper, convex, lower semicontinuous function. In particular, the problem (1.1) corresponds to the special case of $A = 0$, $b = 0$ and (1.6) corresponds to the special case of

$$P(x) = \begin{cases} 0 & \text{if } l \le x \le u; \\ \infty & \text{else.} \end{cases} \tag{3.36}$$

For example, it may be desirable to replace 0 in (3.36) with the 1-norm $\|x\|_1$ to seek a sparse SVM solution. Can the CGD method be extended to solve this more general problem?

Chapter 4

A (BLOCK) COORDINATE GRADIENT DESCENT METHOD FOR LINEARLY CONSTRAINED NONSMOOTH MINIMIZATION

In this chapter, we study the CGD method for solving (1.13). We describe the CGD method formally and we show the global convergence and asymptotic convergence rate of the method. We extend the convergence rate result for the CGD method for solving (1.1) and (1.6) to the problem (1.13) when the coordinate block is chosen by a Gauss-Southwell-type rule. We show that, in the case where P is polyhedral, this rule can be implemented in polynomial time using Rockafellar's notion of conformal realization. This chapter is based on the paper [105] co-authored with P. Tseng.

4.1 (Block) Coordinate Gradient Descent Method

In this section, we describe a (block) coordinate gradient descent method for solving (1.6). In CGD method, we use $\nabla f(x)$ to build a quadratic approximation of f at x and apply coordinate descent to generate an improving feasible direction d at x. More precisely, we choose a nonempty subset $\mathcal{J} \subseteq \mathcal{N}$ and a symmetric matrix $H \in \mathbb{R}^{n\times n}$ (approximating the Hessian $\nabla^2 f(x)$), and move x along the direction $d = d_H(x;\mathcal{J})$, where

$$d_H(x;\mathcal{J}) \stackrel{\rm def}{=} \arg\min_{x+d\in X} \left\{ \nabla f(x)^T d + \frac{1}{2} d^T H d + cP(x+d) - cP(x) \;\mid\; d_j = 0 \;\forall j \notin \mathcal{J} \right\}. \tag{4.1}$$

where $X = \{x \mid l \le x \le u,\ Ax = b\}$. Here $d_H(x;\mathcal{J})$ depends on H through $H_{\mathcal{J}\mathcal{J}}$ only. To ensure that $d_H(x;\mathcal{J})$ is well defined, we assume that $H_{\mathcal{J}\mathcal{J}}$ is positive definite on $\mathrm{Null}(A_{\mathcal{J}})$ or, equivalently, $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succ 0_n$, where $A_{\mathcal{J}}$ denotes the submatrix of

A comprising columns indexed by $\mathcal{J}$ and $B_{\mathcal{J}}$ is a matrix whose columns form an orthonormal basis for Null$(A_{\mathcal{J}})$. The direction (4.1) reduces to those used in Chapter 2 and 3 when $X = \mathbb{R}^n$ or $P \equiv 0$.

Using the convexity of P and the set X, we have the following lemma, analogous to Lemmas 2.1 and 3.1, showing that d is a descent direction at x whenever $d \neq 0$.

Lemma 4.1 *For any $x \in X \cap \mathrm{dom}P$, nonempty $\mathcal{J} \subseteq \mathcal{N}$ and symmetric $H \in \mathbb{R}^{n\times n}$ with $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succ 0_n$, let $d = d_H(x;\mathcal{J})$ and $g = \nabla f(x)$. Then*

$$g^T d + cP(x+d) - cP(x) \leq -d^T H d \leq -\lambda_{\min}(B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}})\|d\|^2. \tag{4.2}$$

Proof. For any $\alpha \in (0,1)$, we have from (4.1) and the convexity of P and X that

$$\begin{aligned} g^T d + \frac{1}{2} d^T H d + cP(x+d) &\leq g^T(\alpha d) + \frac{1}{2}(\alpha d)^T H(\alpha d) + cP(x+\alpha d) \\ &\leq \alpha g^T d + \frac{1}{2}\alpha^2 d^T H d + \alpha cP(x+d) + (1-\alpha)cP(x). \end{aligned}$$

Rearranging terms yields

$$(1-\alpha)g^T d + (1-\alpha)(cP(x+d) - cP(x)) + \frac{1}{2}(1-\alpha^2)d^T H d \leq 0.$$

Since $1-\alpha^2 = (1-\alpha)(1+\alpha)$, dividing both sides by $1-\alpha > 0$ and then taking $\alpha \uparrow 1$ prove the first inequality in (4.2). Since $d_{\mathcal{J}} \in \mathrm{Null}(A_{\mathcal{J}})$ so that $d_{\mathcal{J}} = B_{\mathcal{J}} y$ for some vector y, we have

$$d^T H d = y^T B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} y \geq \|y\|^2 \lambda_{\min}(B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}) = \|d\|^2 \lambda_{\min}(B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}),$$

where the second equality uses $B_{\mathcal{J}}^T B_{\mathcal{J}} = I$. This proves the second inequality in (4.2).

■

We next choose a stepsize $\alpha > 0$ so that $x' = x + \alpha d$ achieves sufficient descent, and re-iterate. We now describe formally the block-coordinate gradient descent method.

CGD method:

Choose $x^0 \in X \cap \mathrm{dom} P$. For $k = 0, 1, 2, ...$, generate x^{k+1} from x^k according to the iteration:

1. Choose a nonempty $\mathcal{J}^k \subseteq \mathcal{N}$ and a symmetric $H^k \in \mathbb{R}^{n\times n}$ with $B_{\mathcal{J}^k}^T H^k_{\mathcal{J}^k\mathcal{J}^k} B_{\mathcal{J}^k} \succ 0_n$.

2. Solve (4.1) with $x = x^k,\ \mathcal{J} = \mathcal{J}^k,\ H = H^k$ to obtain $d^k = d_{H^k}(x^k; \mathcal{J}^k)$.

3. Choose a stepsize $\alpha^k > 0$ and set $x^{k+1} = x^k + \alpha^k d^k$.

Various stepsize rules for smooth optimization [6, 32, 77] can be adapted to our nonsmooth setting to choose α^k. The following adaptation of the Armijo rule, based on Lemma 4.1 and Section 2.1, is simple, requires only function evaluations, and seems effective in theory and practice.

Armijo rule:

Choose $\alpha^k_{\text{init}} > 0$ and let α^k be the largest element of $\{\alpha^k_{\text{init}}\beta^j\}_{j=0,1,...}$ satisfying

$$F_c(x^k + \alpha^k d^k) \le F_c(x^k) + \alpha^k \sigma \Delta^k \quad \text{and} \quad x^k + \alpha^k d^k \in X, \tag{4.3}$$

where $0 < \beta < 1$, $0 < \sigma < 1$, $0 \le \theta < 1$, and

$$\Delta^k \stackrel{\text{def}}{=} \nabla f(x^k)^T d^k + \theta {d^k}^T H^k d^k + cP(x^k + d^k) - cP(x^k). \tag{4.4}$$

Since $B_{\mathcal{J}^k}^T H^k_{\mathcal{J}^k\mathcal{J}^k} B_{\mathcal{J}^k} \succ 0$ and $0 \le \theta < 1$, we see from the convexity of P and X and Lemma 4.1 that

$$\begin{aligned} F_c(x^k + \alpha d^k) &= f(x^k + \alpha d^k) + cP(\alpha(x^k + d^k) + (1-\alpha)x^k) \\ &\le f(x^k + \alpha d^k) + \alpha cP(x^k + d^k) + (1-\alpha)cP(x^k) \\ &= f(x^k) + \alpha \nabla f(x^k)^T d^k + o(\alpha) + \alpha(cP(x^k + d^k) - cP(x^k)) + cP(x^k) \\ &\le F_c(x^k) + \alpha\Delta^k + o(\alpha) \quad \forall \alpha \in (0, 1], \end{aligned}$$

and $\Delta^k \le (\theta - 1){d^k}^T H^k d^k < 0$ whenever $d^k \ne 0$. Since $0 < \sigma < 1$, this shows that α^k given by the Armijo rule is well defined and positive. By choosing α^k_{init} based on

the previous stepsize α^{k-1}, the number of function evaluations can be kept small in practice. Notice that Δ^k increases with θ. Thus, larger stepsizes will be accepted if we choose either σ near 0 or θ near 1.

For convergence, the index subset $\mathcal{J}^k$ must be chosen judiciously. We will choose $\mathcal{J}^k$ according to the *Gauss-Southwell-q* rule, which was introduced in Chapter 2 for the case of $X = \mathbb{R}^n$ and has been shown in Chapter 2 and 3 to be effective in theory and practice. Specifically, let

$$q_H(x;\mathcal{J}) \stackrel{\text{def}}{=} \left\{ \nabla f(x)^T d + \frac{1}{2} d^T H d + cP(x+d) - cP(x) \right\}_{d=d_H(x;\mathcal{J})}, \tag{4.5}$$

which is intuitively the predicted descent when x is moved along the direction $d_H(x;\mathcal{J})$. The Gauss-Southwell-q rule chooses the index subset $\mathcal{J}^k$ to satisfy

$$q_{D^k}(x^k;\mathcal{J}^k) \leq \upsilon \; q_{D^k}(x^k;\mathcal{N}), \tag{4.6}$$

where $D^k \succ 0$ (typically diagonal) and $0 < \upsilon \leq 1$. (In fact, it suffices that $B_{\mathcal{N}}^T D^k B_{\mathcal{N}} \succ 0$ for our analysis.) We will discuss in Section 4.4 how to efficiently implement this rule when P is polyhedral.

4.2 Properties of search direction

In this section we derive various properties of the search direction $d_H(x;\mathcal{J})$ and the corresponding predicted descent $q_H(x;\mathcal{J})$. These properties will be used in later sections to analyze the convergence rate and the complexity of the CGD method.

The following lemma shows that $\|d_H(x;\mathcal{J})\|$ changes not too fast with the quadratic coefficients H. It will be used to prove Theorem 4.2.

Lemma 4.2 *For any $x \in X \cap \mathrm{dom} P$, nonempty $\mathcal{J} \subseteq \mathcal{N}$, and symmetric matrices $H, \tilde{H} \in \mathbb{R}^{n\times n}$ satisfying $U \succ 0_n$ and $\tilde{U} \succ 0_n$, where $U = B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}$ and $\tilde{U} = B_{\mathcal{J}}^T \tilde{H}_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}$. Let $d = d_H(x;\mathcal{J})$ and $\tilde{d} = d_{\tilde{H}}(x;\mathcal{J})$. Then*

$$\|\tilde{d}\| \leq \frac{1 + \lambda_{\max}(S) + \sqrt{1 - 2\lambda_{\min}(S) + \lambda_{\max}(S)^2}}{2} \; \frac{\lambda_{\max}(U)}{\lambda_{\min}(\tilde{U})} \; \|d\|, \tag{4.7}$$

where $S = U^{-1/2}\tilde{U}U^{-1/2}$.

Proof. Since $d_j = \tilde{d}_j = 0$ for all $j \notin \mathcal{J}$, it suffices to prove the lemma for the case of $\mathcal{J} = \mathcal{N}$. Let $g = \nabla f(x)$. By the definition of d and $\tilde{d}$ and applying [90, Theorem 10.1] to (1.14),

$$d \in \arg\min_u \ (g + Hd)^T u + c\hat{P}(x+u) - c\hat{P}(x),$$
$$\tilde{d} \in \arg\min_u \ (g + \tilde{H}\tilde{d})^T u + c\hat{P}(x+u) - c\hat{P}(x),$$

i.e.,

$$d \in \arg\min_u \{(g + Hd)^T u + cP(x+u) - cP(x) \mid x + u \in X\},$$
$$\tilde{d} \in \arg\min_u \{(g + \tilde{H}\tilde{d})^T u + cP(x+u) - cP(x) \mid x + u \in X\}.$$

Thus

$$\begin{aligned}(g + Hd)^T d + cP(x+d) - cP(x) &\leq (g + Hd)^T \tilde{d} + cP(x+\tilde{d}) - cP(x),\\ (g + \tilde{H}\tilde{d})^T \tilde{d} + cP(x+\tilde{d}) - cP(x) &\leq (g + \tilde{H}\tilde{d})^T d + cP(x+d) - cP(x).\end{aligned}$$

Adding the above two inequalities and rearranging terms yield

$$d^T H d - d^T(H + \tilde{H})\tilde{d} + \tilde{d}^T \tilde{H} \tilde{d} \leq 0.$$

Since $d, \tilde{d} \in \text{Null}(A)$, we have $d = B_{\mathcal{N}} y$ and $\tilde{d} = B_{\mathcal{N}} \tilde{y}$ for some vectors $y, \tilde{y}$. Substituting these into the above inequality and using the definitions of $U, \tilde{U}$ yield

$$y^T U y - y^T(U + \tilde{U})\tilde{y} + \tilde{y}^T \tilde{U} \tilde{y} \leq 0.$$

Then proceeding as in the proof of Lemma 2.3 and using $\|d\| = \|y\|$, $\|\tilde{d}\| = \|\tilde{y}\|$ (since $B_{\mathcal{N}}^T B_{\mathcal{N}} = I$), we obtain (4.7). ∎

The next lemma bounds $\nabla f(x)^T(x' - \bar{x}) + cP(x') - cP(\bar{x})$ from above by a weighted sum of $\|x - \bar{x}\|^2$ and $-q_D(x; \mathcal{J})$, where $x' = x + \alpha d$, $d = d_H(x; \mathcal{J})$, and $\mathcal{J}$ satisfies a condition analogous to (4.6). This lemma, which extends Lemma 3.4 for the case of $P \equiv 0$, will be used to prove Theorem 4.2.

Lemma 4.3 *Fix any $x \in X \cap \mathrm{dom}P$, nonempty $\mathcal{J} \subseteq \mathcal{N}$, symmetric matrices $H, D \in \mathbb{R}^{n\times n}$ satisfying $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succ 0_n$, $\bar{\delta} I \succeq D \succ 0_n$, and*

$$q_D(x;\mathcal{J}) \leq \upsilon \; q_D(x;\mathcal{N}), \tag{4.8}$$

with $\bar{\delta} > 0$, $0 < \upsilon \leq 1$. Then, for any $\bar{x} \in X \cap \mathrm{dom}P$, $0 \leq \alpha \leq 1$, we have

$$g^T(x' - \bar{x}) + cP(x') - cP(\bar{x}) \leq \frac{\bar{\delta}}{2}\|\bar{x} - x\|^2 - \frac{1}{\upsilon} q_D(x;\mathcal{J}), \tag{4.9}$$

where $d = d_H(x;\mathcal{J})$, $g = \nabla f(x)$, and $x' = x + \alpha d$.

Proof. Since $\bar{x} - x$ is a feasible solution of the minimization subproblem (4.1) corresponding to $\mathcal{N}$ and D, we have

$$q_D(x;\mathcal{N}) \leq g^T(\bar{x} - x) + \frac{1}{2}(\bar{x} - x)^T D(\bar{x} - x) + cP(\bar{x}) - cP(x).$$

Since $\bar{\delta} I \succeq D \succ 0_n$, we have $0 \leq (\bar{x} - x)^T D(\bar{x} - x) \leq \bar{\delta}\|\bar{x} - x\|^2$. This together with (4.8) yields

$$\frac{1}{\upsilon} q_D(x;\mathcal{J}) \leq g^T(\bar{x} - x) + \frac{\bar{\delta}}{2}\|\bar{x} - x\|^2 + cP(\bar{x}) - cP(x).$$

Rearranging terms, we have

$$g^T(x - \bar{x}) + cP(x) - cP(\bar{x}) \leq \frac{\bar{\delta}}{2}\|\bar{x} - x\|^2 - \frac{1}{\upsilon} q_D(x;\mathcal{J}). \tag{4.10}$$

Also, by the definition of d and (4.2) in Lemma 4.1, for any $\alpha \geq 0$ we have

$$\alpha(g^T d + cP(x + d) - cP(x)) \leq 0.$$

Since P is convex so that $cP(x + \alpha d) - cP(x) \leq \alpha(cP(x + d) - cP(x))$, this implies

$$\alpha g^T d + cP(x + \alpha d) - cP(x) \leq 0.$$

Adding this to (4.10) yields (4.9). ∎

The next lemma shows that Δ is bounded above by a constant multiple of $q_H(x;\mathcal{J})$. It also bounds $q_H(x;\mathcal{J})$ from above by a constant multiple of $q_D(x;\mathcal{J})$. This lemma will be used to prove Theorem 4.3.

Lemma 4.4 *For any $x \in X \cap \mathrm{dom}P$, nonempty $\mathcal{J} \subseteq \mathcal{N}$, and symmetric matrix $H \in \mathbb{R}^{n\times n}$ satisfying $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succ 0$, the following results hold with $d = d_H(x;\mathcal{J})$ and $g = \nabla f(x)$.*

(a) *For any $0 \le \theta < 1$,*

$$\Delta \le \min\{1, 2-2\theta\} q_H(x;\mathcal{J}),$$

where $\Delta = g^T d + \theta d^T H d + cP(x+d) - cP(x)$.

(b) *For any symmetric matrix $D \in \mathbb{R}^{n\times n}$ satisfying $B_{\mathcal{J}}^T D_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succeq B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}$ and any $0 < \omega \le 1$,*

$$q_H(x;\mathcal{J}) \le q_D(x;\mathcal{J}) \le \omega q_{\omega D}(x;\mathcal{J}).$$

Proof. (a) If $\theta \le 1/2$, then $d^T H d \le 0$ by (4.2) in Lemma 2.1, so that

$$\Delta = q_H(x;\mathcal{J}) + (\theta - \frac{1}{2}) d^T H d \le q_H(x;\mathcal{J}).$$

Otherwise, $1/2 < \theta < 1$ and we have from (4.2) in Lemma 2.1 that

$$\begin{aligned}
& \Delta \\
= \; & g^T d + (2\theta - 1) d^T H d + (1-\theta) d^T H d + cP(x+d) - cP(x) \\
\le \; & g^T d + (2\theta - 1)(-g^T d - cP(x+d) + cP(x)) \\
& + (1-\theta) d^T H d + cP(x+d) - cP(x) \\
= \; & (2 - 2\theta) q_H(x;\mathcal{J}).
\end{aligned}$$

Thus $\Delta \le \min\{1, 2-2\theta\} q_H(x;\mathcal{J})$.

(b) Let $\bar{d} = d_D(x;\mathcal{J})$. Then

$$\begin{aligned}
q_H(x;\mathcal{J}) &= g^T d + \frac{1}{2} d^T H d + cP(x+d) - cP(x) \\
&\le g^T \bar{d} + \frac{1}{2} \bar{d}^T H \bar{d} + cP(x+\bar{d}) - cP(x) \\
&\le g^T \bar{d} + \frac{1}{2} \bar{d}^T D \bar{d} + cP(x+\bar{d}) - cP(x) \\
&= q_D(x;\mathcal{J}),
\end{aligned}$$

where the third step uses $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \preceq B_{\mathcal{J}}^T D_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}$. This proves the first inequality. To prove the second inequality, we note that, by using (1.14),

$$\begin{aligned}
& q_{\omega D}(x;\mathcal{J}) \\
= & \min_{u_j=0 \ \forall j \notin \mathcal{J}} \left\{ g^T u + \frac{\omega}{2} u^T D u + c\hat{P}(x+u) - c\hat{P}(x) \right\} \\
= & \frac{1}{\omega} \min_{u_j=0 \ \forall j \notin \mathcal{J}} \left\{ g^T(\omega u) + \frac{1}{2}(\omega u)^T D(\omega u) + \omega(c\hat{P}(x+u) - c\hat{P}(x)) \right\} \\
\geq & \frac{1}{\omega} \min_{u_j=0 \ \forall j \notin \mathcal{J}} \left\{ g^T(\omega u) + \frac{1}{2}(\omega u)^T D(\omega u) + c\hat{P}(x+\omega u) - c\hat{P}(x) \right\} \\
= & \frac{1}{\omega} q_D(x;\mathcal{J}),
\end{aligned}$$

where the inequality uses the convexity of $\hat{P}$. ■

Corollary 4.1 *For any $x \in X \cap \mathrm{dom} P$, nonempty $\mathcal{J} \subseteq \mathcal{N}$, and symmetric matrices $H, D \in \mathbb{R}^{n\times n}$ satisfying $0_n \prec B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \preceq \bar{\lambda} I$ and $B_{\mathcal{J}}^T D_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succeq \underline{\delta} I$, we have*

$$q_H(x;\mathcal{J}) \leq \min\left\{1, \frac{\underline{\delta}}{\bar{\lambda}}\right\} q_D(x;\mathcal{J}).$$

Proof. By assumption on $H_{\mathcal{J}\mathcal{J}}$ and $D_{\mathcal{J}\mathcal{J}}$, we have $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \preceq \frac{\bar{\lambda}}{\underline{\delta}} B_{\mathcal{J}}^T D_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}$. If $\frac{\bar{\lambda}}{\underline{\delta}} \leq 1$, then $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \preceq \frac{\bar{\lambda}}{\underline{\delta}} B_{\mathcal{J}}^T D_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \preceq B_{\mathcal{J}}^T D_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}$, so Lemma 4.4(b) yields $q_H(x;\mathcal{J}) \leq q_D(x;\mathcal{J})$. If $\frac{\bar{\lambda}}{\underline{\delta}} > 1$, then Lemma 4.4(b) again yields

$$q_H(x;\mathcal{J}) \leq q_{\frac{\bar{\lambda}}{\underline{\delta}} D}(x;\mathcal{J}) \leq \frac{\underline{\delta}}{\bar{\lambda}} q_D(x;\mathcal{J}).$$

This proves the desired result. ■

4.3 Convergence Rate Analysis

In this section we analyze the global convergence and asymptotic convergence rate of the CGD method using the Gauss-Southwell-q rule, analogous to Theorems 2.1, 2.4 for the case $X = \mathbb{R}^n$, and 3.2 for the case of $P \equiv 0$. Analogous to Chapter 3, we make the following assumption on $\{H^k\}$ in the CGD method.

Assumption 4.1 *$\bar{\lambda} I \succeq B_{\mathcal{J}^k}^T H_{\mathcal{J}^k \mathcal{J}^k}^k B_{\mathcal{J}^k} \succeq \underline{\lambda} I$ for all k, where $0 < \underline{\lambda} \leq \bar{\lambda}$.*

Theorem 4.1 *Let $\{x^k\}$, $\{\mathcal{J}^k\}$, $\{H^k\}$, $\{d^k\}$ be sequences generated by the CGD method under Assumption 4.1, where $\{\alpha^k\}$ is chosen by the Armijo rule with $\inf_k \alpha_{\text{init}}^k > 0$. Then the following results hold.*

(a) *$\{F_c(x^k)\}$ is nonincreasing and Δ^k given by (4.4) satisfies*

$$-\Delta^k \geq (1-\theta) {d^k}^T H^k d^k \geq (1-\theta)\underline{\lambda} \|d^k\|^2 \quad \forall k, \tag{4.11}$$

$$F_c(x^{k+1}) - F_c(x^k) \leq \sigma \alpha^k \Delta^k \leq 0 \quad \forall k. \tag{4.12}$$

(b) *If $\{\mathcal{J}^k\}$ satisfies (4.6), $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k, where $0 < \underline{\delta} \leq \bar{\delta}$, and either (1) P is continuous on $X \cap \text{dom} P$ or (2) $\inf_k \alpha^k > 0$ or (3) $\alpha_{\text{init}}^k = 1$ for all k, then every cluster point of $\{x^k\}$ is a stationary point of F_c.*

(c) *If f satisfies*

$$\|\nabla f(y) - \nabla f(z)\| \leq L \|y - z\| \quad \forall y,\, z \in X \cap \text{dom} P, \tag{4.13}$$

for some $L \geq 0$, then $\alpha^k \geq \min\{\alpha_{\text{init}}^k, \beta \min\{1, 2\underline{\lambda}(1-\sigma+\sigma\theta)/L\}$ for all k. If $\lim_{k\to\infty} F_c(x^k) > -\infty$ also, then $\{\Delta^k\} \to 0$ and $\{d^k\} \to 0$.

Proof. The proof is nearly identical to that of Theorem 2.1(a), (b), (d), (f), applied to the problem (1.14). Assumption 4.1 is weaker than Assumption 2.1, but it suffices. ■

Since P is separable, P is automatically continuous on $\text{dom} P$ [90, Corollary 2.37]. The next theorem establishes the convergence rate of the CGD method under Assumption 4.1 and the following assumption that is analogous to Assumption 2.2. In what follows, $\bar{X}$ denotes the set of stationary points of F_c and

$$\text{dist}(x, \bar{X}) = \min_{\bar{x} \in \bar{X}} \|x - \bar{x}\| \qquad \forall x \in \mathbb{R}^n.$$

Assumption 4.2 **(a)** *$\bar{X} \neq \emptyset$ and, for any $\zeta \geq \min_x F_c(x)$, there exist scalars $\tau > 0$ and $\epsilon > 0$ such that*

$$\mathrm{dist}(x, \bar{X}) \leq \tau \|d_I(x; \mathcal{N})\| \quad \textit{whenever } x \in X,\ F_c(x) \leq \zeta,\ \|d_I(x; \mathcal{N})\| \leq \epsilon.$$

(b) *There exists a scalar $\rho > 0$ such that*

$$\|x - y\| \geq \rho \quad \textit{whenever} \quad x \in \bar{X},\ y \in \bar{X},\ F_c(x) \neq F_c(y).$$

Assumption 4.2(a) is a local Lipschitzian error bound assumption, saying that the distance from x to $\bar{X}$ is locally in the order of the norm of the residual at x. Assumption 4.2(b) says that the isocost surfaces of F_c restricted to the solution set $\bar{X}$ are "properly separated." Assumption 4.2(b) holds automatically if f is convex or f is quadratic and P is polyhedral; see Section 2.5 for further discussions. Upon applying Theorem 2.5 to the problem (1.14), we obtain the following sufficient conditions for Assumption 4.2(a) to hold.

Proposition 4.1 *Suppose that $\bar{X} \neq \emptyset$ and any of the following conditions hold.*

C1 *f is strongly convex and satisfies* (4.13) *for some $L \geq 0$.*

C2 *f is quadratic. P is polyhedral.*

C3 *$f(x) = g(Ex) + q^T x$ for all $x \in \mathbb{R}^n$, where $E \in \mathbb{R}^{m \times n}$, $q \in \mathbb{R}^n$, and g is a strongly convex differentiable function on $\mathbb{R}^m$ with ∇g Lipschitz continuous on $\mathbb{R}^m$. P is polyhedral.*

C4 *$f(x) = \max_{y \in Y}\{(Ex)^T y - g(y)\} + q^T x$ for all $x \in \mathbb{R}^n$, where Y is a polyhedral set in $\mathbb{R}^m$, $E \in \mathbb{R}^{m \times n}$, $q \in \mathbb{R}^n$, and g is a strongly convex differentiable function on $\mathbb{R}^m$ with ∇g Lipschitz continuous on $\mathbb{R}^m$. P is polyhedral.*

Then Assumption 4.2(a) holds.

The next theorem establishes, under Assumptions 4.1 and 4.2, the linear rate of convergence of the CGD method using (4.6) to choose $\{\mathcal{J}^k\}$. Its proof, based on the ideas in Theorem 3.2 for the case of $P \equiv 0$, uses Theorem 4.1 and Lemmas 4.1, 4.2, 4.3.

Theorem 4.2 *Assume that f satisfies (4.13) for some $L \geq 0$. Let $\{x^k\}$, $\{H^k\}$, $\{d^k\}$ be sequences generated by the CGD method satisfying Assumption 4.1, where $\{\mathcal{J}^k\}$ is chosen by (4.6) with $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k $(0 < \underline{\delta} \leq \bar{\delta})$. If F_c satisfies Assumption 4.2 and $\{\alpha^k\}$ is chosen by the Armijo rule with $\sup_k \alpha^k_{\rm init} \leq 1$ and $\inf_k \alpha^k_{\rm init} > 0$, then either $\{F_c(x^k)\} \downarrow -\infty$ or $\{F_c(x^k)\}$ converges at least Q-linearly and $\{x^k\}$ converges at least R-linearly to a point in $\bar{X}$.*

Proof. For each $k = 0, 1, \ldots$, (4.4) and $d^k = d_{H^k}(x^k; \mathcal{J}^k)$ imply that

$$\begin{aligned} \Delta^k + \left(\frac{1}{2} - \theta\right) {d^k}^T H^k d^k &= {g^k}^T d^k + \frac{1}{2} {d^k}^T H^k d^k + cP(x^k + d^k) - cP(x^k) \\ &\leq {g^k}^T \tilde{d}^k + \frac{1}{2} (\tilde{d}^k)^T H^k \tilde{d}^k + cP(x^k + \tilde{d}^k) - cP(x^k) \\ &= q_{D^k}(x^k; \mathcal{J}^k) + \frac{1}{2} (\tilde{d}^k)^T (H^k - D^k) \tilde{d}^k, \end{aligned} \tag{4.14}$$

where we let $\tilde{d}^k = d_{D^k}(x^k; \mathcal{J}^k)$. By Lemma 4.2 with $\mathcal{J} = \mathcal{J}^k$, $H = H^k$ and $\tilde{H} = D^k$,

$$\|d_{D^k}(x^k; \mathcal{J}^k)\| \leq \frac{1 + \bar{\delta}/\underline{\lambda} + \sqrt{1 - 2\underline{\delta}/\bar{\lambda} + (\bar{\delta}/\underline{\lambda})^2}}{2} \, \frac{\bar{\lambda}}{\underline{\delta}} \, \|d^k\|. \tag{4.15}$$

This together with (4.14) and $(\tilde{d}^k)^T (H^k - D^k) \tilde{d}^k \leq (\bar{\lambda} - \underline{\delta}) \|\tilde{d}^k\|^2$ implies that

$$\Delta^k + \left(\frac{1}{2} - \theta\right) {d^k}^T H^k d^k \leq q_{D^k}(x^k; \mathcal{J}^k) + \omega \|d^k\|^2. \tag{4.16}$$

Here, $\omega \in \mathbb{R}$ is a constant depending on $\bar{\lambda}, \underline{\lambda}, \bar{\delta}, \underline{\delta}$ only. Also, by (4.5) and Lemma 4.1 with $\mathcal{J} = \mathcal{N}$, $H = D^k$, we have

$$\begin{aligned} q_{D^k}(x^k; \mathcal{N}) &= \left((g^k)^T d + \frac{1}{2} d^T D^k d + cP(x^k + d) - cP(x^k) \right)_{d = d_{D^k}(x^k; \mathcal{N})} \\ &\leq \left(-\frac{1}{2} d^T D^k d \right)_{d = d_{D^k}(x^k; \mathcal{N})} \\ &\leq -\frac{\underline{\delta}}{2} \|d_{D^k}(x^k; \mathcal{N})\|^2 \quad \forall k, \end{aligned} \tag{4.17}$$

where the last inequality uses $D^k \succeq \underline{\delta} I$.

By Theorem 4.1(a), $\{F_c(x^k)\}$ is nonincreasing. Thus either $\{F_c(x^k)\} \downarrow -\infty$ or $\lim_{k\to\infty} F_c(x^k) > -\infty$. Suppose the latter. Since α^k is chosen by the Armijo rule with $\inf_k \alpha^k_{\text{init}} > 0$, Theorem 4.1(c) implies $\inf_k \alpha^k > 0$, $\{\Delta^k\} \to 0$, and $\{d^k\} \to 0$. Since $\{H^k\}$ is bounded by Assumption 4.1, we obtain from (4.16) that $0 \leq \lim_{k\to\infty} \inf q_{D^k}(x^k; \mathcal{J}^k)$. Then (4.6) and (4.17) yield $\{d_{D^k}(x^k; \mathcal{N})\} \to 0$.

By Lemma 4.2 with $\mathcal{J} = \mathcal{N}$, $H = D^k$ and $\tilde{H} = I$, we have

$$\|d_I(x^k; \mathcal{N})\| \leq \frac{1 + 1/\underline{\delta} + \sqrt{1 - 2/\bar{\delta} + (1/\underline{\delta})^2}}{2} \bar{\delta} \, \|d_{D^k}(x^k; \mathcal{N})\| \quad \forall k. \tag{4.18}$$

Hence $\{d_I(x^k; \mathcal{N})\} \to 0$. Since $\{F_c(x^k)\}$ is nonincreasing, this implies that $F_c(x^k) \leq F_c(x^0)$ and $\|d_I(x^k; \mathcal{N})\| \leq \epsilon$ for all $k \geq$ some $\bar{k}$. Then, by Assumption 4.2(a), there exist $\bar{k}$ and $\tau > 0$ such that

$$\|x^k - \bar{x}^k\| \leq \tau \|d_I(x^k; \mathcal{N})\| \quad \forall k \geq \bar{k}, \tag{4.19}$$

where $\bar{x}^k \in \bar{X}$ satisfies $\|x^k - \bar{x}^k\| = \text{dist}(x^k, \bar{X})$. Since $\{d_I(x^k; \mathcal{N})\} \to 0$, this implies $\{x^k - \bar{x}^k\} \to 0$. Since $\{x^{k+1} - x^k\} = \{\alpha^k d^k\} \to 0$, this and Assumption 4.2(b) imply that $\{\bar{x}^k\}$ eventually settles down at some isocost surface of F_c, i.e., there exist an index $\hat{k} \geq \bar{k}$ and a scalar $\bar{v}$ such that

$$F_c(\bar{x}^k) = \bar{v} \quad \forall k \geq \hat{k}. \tag{4.20}$$

Fix any index $k \geq \hat{k}$. Since $\bar{x}^k$ is a stationary point of F_c, we have

$$\nabla f(\bar{x}^k)^T (x^k - \bar{x}^k) + cP(x^k) - cP(\bar{x}^k) \geq 0.$$

We also have from the Mean Value Theorem that

$$f(x^k) - f(\bar{x}^k) = \nabla f(\psi^k)^T (x^k - \bar{x}^k),$$

for some ψ^k lying on the line segment joining x^k with $\bar{x}^k$. Since $x^k, \bar{x}^k$ lie in the convex set $X \cap \text{dom} P$, so does ψ^k. Combining these two relations and using (4.20), we obtain

$$\bar{v} - F_c(x^k) \quad \leq \quad (\nabla f(\bar{x}^k) - \nabla f(\psi^k))^T (x^k - \bar{x}^k)$$

$$\begin{aligned} &\leq \|\nabla f(\bar{x}^k) - \nabla f(\psi^k)\|\|x^k - \bar{x}^k\| \\ &\leq L\|x^k - \bar{x}^k\|^2, \end{aligned}$$

where the last inequality uses (4.13), the convexity of $X \cap \mathrm{dom}P$, and $\|\psi^k - \bar{x}^k\| \leq \|x^k - \bar{x}^k\|$. This together with $\{x^k - \bar{x}^k\} \to 0$ proves that

$$\liminf_{k\to\infty} F_c(x^k) \geq \bar{v}. \tag{4.21}$$

For each index $k \geq \hat{k}$, we have from (4.20) that

$$\begin{aligned} & F_c(x^{k+1}) - \bar{v} \\ = \; & f(x^{k+1}) + cP(x^{k+1}) - f(\bar{x}^k) - cP(\bar{x}^k) \\ = \; & \nabla f(\tilde{x}^k)^T(x^{k+1} - \bar{x}^k) + cP(x^{k+1}) - cP(\bar{x}^k) \\ = \; & (\nabla f(\tilde{x}^k) - \nabla f(x^k))^T(x^{k+1} - \bar{x}^k) + \nabla f(x^k)^T(x^{k+1} - \bar{x}^k) + cP(x^{k+1}) - cP(\bar{x}^k) \\ \leq \; & L\|\tilde{x}^k - x^k\|\|x^{k+1} - \bar{x}^k\| + \frac{\bar{\delta}}{2}\|x^k - \bar{x}^k\|^2 - \frac{1}{\upsilon} q_{D^k}(x^k; \mathcal{J}^k), \end{aligned} \tag{4.22}$$

where the second step uses the Mean Value Theorem with $\tilde{x}^k$ a point lying on the segment joining x^{k+1} with $\bar{x}^k$ (so that $\tilde{x}^k \in X$); the fourth step uses (4.13) and Lemma 4.3. Using the inequalities $\|\tilde{x}^k - x^k\| \leq \|x^{k+1} - x^k\| + \|x^k - \bar{x}^k\|$, $\|x^{k+1} - \bar{x}^k\| \leq \|x^{k+1} - x^k\| + \|x^k - \bar{x}^k\|$ and $\|x^{k+1} - x^k\| = \alpha^k\|d^k\|$, we see from (4.19), and $\sup_k \alpha^k \leq 1$ (since $\sup_k \alpha^k_{\mathrm{init}} \leq 1$) that the right-hand side of (4.22) is bounded above by

$$C_1\left(\|d^k\|^2 - q_{D^k}(x^k; \mathcal{J}^k) + \|d_I(x^k; \mathcal{N})\|^2\right) \tag{4.23}$$

for all $k \geq \hat{k}$, where $C_1 > 0$ is some constant depending on $L, \tau, \bar{\delta}, \upsilon$ only.

By (4.11), we have

$$\underline{\lambda}\|d^k\|^2 \leq {d^k}^T H^k d^k \leq -\frac{1}{1-\theta}\Delta^k \quad \forall k. \tag{4.24}$$

By (4.17) and (4.18), we also have

$$\|d_I(x^k; \mathcal{N})\|^2 \leq \left(1 + 1/\underline{\delta} + \sqrt{1 - 2/\bar{\delta} + (1/\underline{\delta})^2}\right)^2 \frac{\bar{\delta}^2}{2\underline{\delta}}\left(-q_{D^k}(x^k; \mathcal{N})\right) \quad \forall k.$$

Thus, the quantity in (4.23) is bounded above by

$$C_2\left(-\Delta^k - q_{D^k}(x^k;\mathcal{J}^k) - q_{D^k}(x^k;\mathcal{N})\right) \tag{4.25}$$

for all $k \geq \hat{k}$, where $C_2 > 0$ is some constant depending on $L, \tau, \bar{\delta}, \underline{\delta}, \theta, \underline{\lambda}, \upsilon$ only.

Combining (4.16) with (4.24) yields

$$\begin{aligned} -q_{D^k}(x^k;\mathcal{J}^k) &\leq -\Delta^k + \left(\theta - \frac{1}{2}\right) d^{kT} H^k d^k + \omega\|d^k\|^2 \\ &\leq -\Delta^k - \max\left\{0, \theta - \frac{1}{2}\right\} \frac{1}{1-\theta}\Delta^k - \frac{\omega}{\underline{\lambda}(1-\theta)}\Delta^k. \end{aligned} \tag{4.26}$$

Combining (4.6) and (4.26), we see that the quantity in (4.25) is bounded above by

$$-C_3\Delta^k$$

all $k \geq \hat{k}$, where $C_3 > 0$ is some constant depending on $L, \tau, \bar{\delta}, \underline{\delta}, \theta, \bar{\lambda}, \underline{\lambda}, \upsilon$ only. Thus the right-hand side of (4.22) is bounded above by $-C_3\Delta^k$ for all $k \geq \hat{k}$. Combining this with (4.12), (4.22), and $\inf_k \alpha^k > 0$ (see Theorem 4.1(c)) yields

$$F_c(x^{k+1}) - \bar{\upsilon} \leq C_4(F_c(x^k) - F_c(x^{k+1})) \quad \forall k \geq \hat{k},$$

where $C_4 = C_3/(\sigma \inf_k \alpha^k)$. Upon rearranging terms and using (4.21), we have

$$0 \leq F_c(x^{k+1}) - \bar{\upsilon} \leq \frac{C_4}{1+C_4}(F_c(x^k) - \bar{\upsilon}) \quad \forall k \geq \hat{k},$$

so $\{F_c(x^k)\}$ converges to $\bar{\upsilon}$ at least Q-linearly.

Finally, by (4.12), (4.24), and $x^{k+1} - x^k = \alpha^k d^k$, we have

$$\sigma(1-\theta)\underline{\lambda}\frac{\|x^{k+1} - x^k\|^2}{\alpha^k} \leq F_c(x^k) - F_c(x^{k+1}) \quad \forall k \geq \hat{k}.$$

This implies

$$\|x^{k+1} - x^k\| \leq \sqrt{\frac{\sup_k \alpha^k}{\sigma(1-\theta)\underline{\lambda}}(F_c(x^k) - F_c(x^{k+1}))} \quad \forall k \geq \hat{k}.$$

Since $\{F_c(x^k) - F_c(x^{k+1})\} \to 0$ at least R-linearly and $\sup_k \alpha^k \leq 1$, this implies that $\{x^k\}$ converges at least R-linearly. ∎

The assumption (4.13) in Theorem 4.2 can be relaxed to ∇f being Lipschitz continuous on $X \cap \mathrm{dom}P \cap (X^0 + \varrho B)$ for some $\varrho > 0$, where B denotes the unit Euclidean ball in $\mathbb{R}^n$ and X^0 denotes the convex hull of the level set $\{x \mid F_c(x) \leq F_c(x^0)\}$. For simplicity, we did not consider this more relaxed assumption.

In the proof of Theorem 4.2, we do not use the separability assumption on P. Hence Theorem 4.2 can be extended to the case where P is nonseparable. Therefore Theorem 4.2 is an extension of Theorem 2.4 in which we assume that P is block-separable.

4.4 Complexity analysis when f is convex

The following theorem is the main result of this section, giving an upper bound on the number of iterations for the CGD method to achieve ϵ-optimality when f is convex with Lipschitz continuous gradient. Its proof uses Lemmas 4.1, 4.4, and Theorem 4.1(c).

Theorem 4.3 *Suppose f is convex and satisfies* (4.13) *for some $L \geq 0$. Let $\{x^k\}$, $\{\mathcal{J}^k\}$, $\{H^k\}$ be sequences generated by the CGD method under Assumption 4.1, where $\{\mathcal{J}^k\}$ satisfies* (4.6) *with $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k ($0 < \underline{\delta} \leq \bar{\delta}$), and $\{\alpha^k\}$ is chosen by the Armijo rule with $\inf_k \alpha^k_{\mathrm{init}} > 0$. Let $e^k = F_c(x^k) - \min_{x \in X} F_c(x)$ for all k. Then $e^k \leq \epsilon$ whenever*

$$k \geq \begin{cases} \max\left\{0, \left\lfloor \frac{2}{C\sigma\underline{\alpha}} \ln\left(\frac{e^0}{\epsilon}\right) \right\rfloor\right\} & \text{if } \epsilon > \bar{\delta} r^0; \\ \left\lfloor \frac{\bar{\delta} r^0}{C\sigma\underline{\alpha}\epsilon} \right\rfloor + \max\left\{0, \left\lfloor \frac{2}{C\sigma\underline{\alpha}} \ln\left(\frac{e^0}{\bar{\delta} r^0}\right) \right\rfloor\right\} + 1 & \text{else,} \end{cases}$$

where $r^0 = \max_{x \in X}\left\{\mathrm{dist}(x, \bar{X})^2 \mid F_c(x) \leq F_c(x^0)\right\}$, $\bar{X} = \arg\min_{x \in X} F_c(x)$, $C = \min\{1, 2 - 2\theta\} \min\{1, \underline{\delta}/\bar{\lambda}\}\upsilon$, and $\underline{\alpha} = \min\{\inf_k \alpha^k_{\mathrm{init}}, \beta \min\{1, 2\underline{\lambda}(1 - \sigma + \sigma\theta)/L\}$.

Proof. For each $k = 0, 1, \ldots$, by (4.3), (4.6), and Corollary 4.1 with $H = H^k$ and $D = D^k$, and Lemma 4.4(a), we have

$$e^{k+1} - e^k = F_c(x^{k+1}) - F_c(x^k) \leq C\sigma\alpha^k q_{D^k}(x^k; \mathcal{N}). \tag{4.27}$$

For each $k = 0, 1, \ldots,$ and $t \in [0,1]$, let $g^k = \nabla f(x^k)$ and let $\bar{x}^k \in \bar{X}$ satisfy $\|x^k - \bar{x}^k\| = \mathrm{dist}(x^k, \bar{X})$. Then, by using (1.14),

$$\begin{aligned}
q_{D^k}(x^k;\mathcal{N}) &= \min_{d\in\mathbb{R}^n} g^{kT}d + \frac{1}{2}d^T D^k d + c\hat{P}(x^k+d) - c\hat{P}(x^k) \\
&\leq g^{kT}t(\bar{x}^k - x^k) + \frac{t^2}{2}(\bar{x}^k - x^k)^T D^k(\bar{x}^k - x^k) \\
&\quad + c\hat{P}(x^k + t(\bar{x}^k - x^k)) - c\hat{P}(x^k) \\
&\leq g^{kT}t(\bar{x}^k - x^k) + \frac{t^2}{2}(\bar{x}^k - x^k)^T D^k(\bar{x}^k - x^k) + tc\hat{P}(\bar{x}^k) - tc\hat{P}(x^k) \\
&\leq t(f(\bar{x}^k) - f(x^k)) + tc\hat{P}(\bar{x}^k) - tc\hat{P}(x^k) + \frac{t^2}{2}\bar{\delta}\mathrm{dist}(x^k, \bar{X})^2 \\
&= -te^k + \frac{t^2}{2}\bar{\delta}\mathrm{dist}(x^k, \bar{X})^2 \\
&\leq -te^k + \frac{t^2}{2}\bar{\delta}r^0.
\end{aligned}$$

where the second inequality uses the convexity of $\hat{P}$ and the third inequality uses the convexity of f. This holds for all $t \in [0,1]$. Minimizing the right-hand side with respect to t yields

$$q_{D^k}(x^k;\mathcal{N}) \leq -\frac{(e^k)^2}{2\bar{\delta}r^0}$$

if $e^k \leq \bar{\delta}r^0$; and else

$$q_{D^k}(x^k;\mathcal{N}) \leq -e^k + \frac{1}{2}\bar{\delta}r^0 < -\frac{1}{2}e^k.$$

This together with (4.27) yields that

$$e^{k+1} \leq e^k - C\begin{cases} \frac{\sigma\alpha^k}{\bar{\delta}r^0}(e^k)^2 & \text{if } e^k \leq \bar{\delta}r^0; \\ \frac{\sigma\alpha^k}{2}e^k & \text{else.} \end{cases}$$

By Theorem 4.1(c), $\alpha^k \geq \underline{\alpha}$. Hence

$$e^{k+1} \leq e^k - C\frac{\sigma\underline{\alpha}}{\bar{\delta}r^0}(e^k)^2 = e^k\left(1 - C\frac{\sigma\underline{\alpha}}{\bar{\delta}r^0}(e^k)\right) \tag{4.28}$$

if $e^k \leq \bar{\delta}r^0$; and else

$$e^{k+1} \leq e^k - C\frac{\sigma\underline{\alpha}}{2}e^k. \tag{4.29}$$

Case (1): If $\epsilon > \bar{\delta} r^0$, then (4.29) implies $e^k \leq \epsilon$ whenever

$$e^0 \left(1 - C\frac{\sigma\underline{\alpha}}{2}\right)^k < e^0 \exp(-kC\sigma\underline{\alpha}/2) \leq \epsilon$$

or, equivalently,

$$k \;\geq\; \max\left\{0, \left\lfloor \frac{2}{C\sigma\underline{\alpha}} \ln\left(\frac{e^0}{\epsilon}\right) \right\rfloor\right\}.$$

Here and in what follows, $\lfloor \cdot \rfloor$ is the floor function.

Case (2): If $\epsilon \leq \bar{\delta} r^0$, then (4.29) implies $e^k \leq \bar{\delta} r^0$ whenever

$$e^0 \left(1 - C\frac{\sigma\underline{\alpha}}{2}\right)^k < e^0 \exp(-k_0 C\sigma\underline{\alpha}/2) \leq \bar{\delta} r^0$$

or, equivalently,

$$k \;\geq\; k_0 \stackrel{\text{def}}{=} \max\left\{0, \left\lfloor \frac{2}{C\sigma\underline{\alpha}} \ln\left(\frac{e^0}{\bar{\delta} r^0}\right) \right\rfloor\right\}.$$

For each $k \geq k_0$, $e^k \leq \bar{\delta} r^0$. If $e^k = 0$, then $e^k \leq \epsilon$. Otherwise $e^k > 0$. Then $e^j > 0$ for $j = 0, 1, \ldots, k$ and we consider the reciprocals $\xi_j = 1/e^j$. By (4.28) and $e^k > 0$, we have $0 \leq C_1 e^j < 1$ for $j = 0, 1, \ldots, k-1$, where $C_1 = C\frac{\sigma\underline{\alpha}}{\bar{\delta} r^0}$. Thus (4.28) yields

$$\xi_{j+1} - \xi_j \geq \frac{1}{e^j(1 - C_1 e^j)} - \frac{1}{e^j} = \frac{C_1}{1 - C_1 e^j} \geq C_1, \quad j = 0, 1, \ldots, k-1.$$

Therefore $\xi_k = \xi_{k_0} + \sum_{j=k_0}^{k-1} (\xi_{j+1} - \xi_j) \geq C_1(k - k_0)$ and consequently

$$e^k = \frac{1}{\xi_k} \leq \frac{1}{C_1(k - k_0)}.$$

It follows that $e^k \leq \epsilon$ whenever

$$k \geq k_0 + \left\lfloor \frac{1}{C_1\epsilon} \right\rfloor + 1 = \left\lfloor \frac{\bar{\delta} r^0}{C\sigma\underline{\alpha}\epsilon} \right\rfloor + \max\left\{0, \left\lfloor \frac{2}{C\sigma\underline{\alpha}} \ln\left(\frac{e^0}{\bar{\delta} r^0}\right) \right\rfloor\right\} + 1.$$

■

If we take $\theta = 1/2$, $D^k = H^k = I$ and $\alpha_{\text{init}}^k = 1$ for all k, then $\underline{\delta} = \bar{\delta} = \underline{\lambda} = \bar{\lambda}$ and $C = \upsilon$, and the iteration bounds in Theorem 4.3 reduce to

$$\begin{cases} O\left(\frac{L}{\upsilon}\max\left\{0, \ln\left(\frac{\epsilon^0}{\epsilon}\right)\right\}\right) & \text{if } \epsilon > r^0; \\ O\left(\frac{Lr^0}{\upsilon\epsilon} + \frac{L}{\upsilon}\max\left\{0, \ln\left(\frac{\epsilon^0}{r^0}\right)\right\}\right) & \text{else.} \end{cases}$$

Notice that $r^0 = 0$ whenever $x^0 \in \bar{X}$. If $\bar{X}$ is bounded, then it can be seen that $r^0 \to 0$ as $\text{dist}(x^0, \bar{X}) \to 0$.

4.5 Index Subset Selection

In this section we study efficient ways to find an index subset $\mathcal{J}^k$ satisfying (4.6) for some constant $0 < \upsilon \leq 1$. One obvious choice is $\mathcal{J}^k = \mathcal{N}$, which satisfies (4.6) with $\upsilon = 1$. However, the corresponding search direction (4.1) may be expensive to compute and, for SVM applications, the gradient would be expensive to update. We will extend the procedure developed in Chapter 2, involving a conformal realization of $d_{D^k}(x^k; \mathcal{N})$ [87], [89, Section 10B], to find $\mathcal{J}^k$ of small size for the case where P is separable. Our main result is Proposition 4.2, showing the existence of such $\mathcal{J}^k$ by construction.

First, we derive a lower bound on $P(x+d) - P(x)$, based on a conformal realization of d. This bound will be used to prove Proposition 4.2.

Lemma 4.5 *For any $x, x+d \in X \cap \text{dom}P$, let d be expressed as $d = d^1 + \cdots + d^r$, for some $r \geq 1$ and some nonzero $d^t \in \text{Null}(A)$ conformal to d with $t = 1, \ldots, r$. Then*

$$P(x+d) - P(x) \geq \sum_{t=1}^{r} \left(P(x+d^t) - P(x)\right).$$

Proof. Since P is separable, it suffices to prove that, for $j \in \mathcal{N}$,

$$P_j(x_j + d_j^1 + \cdots + d_j^r) - P_j(x_j) \geq \sum_{t=1}^{r} \left(P_j(x_j + d_j^t) - P_j(x_j)\right). \tag{4.30}$$

We prove this by induction on r. This clearly holds for $r = 1$. Suppose (4.5) holds for $r < s$, where $s \geq 2$. We show below that (4.30) holds for $r = s$. If $d_j^1 + \cdots + d_j^{s-1} = 0$,

then (4.30) reduces to the case of $r = 1$ and hence holds. If $d_j^s = 0$, then (4.30) reduces to the case of $r < s$ and hence holds. Thus it remains to consider the case of $d_j^1 + \cdots + d_j^{s-1} \neq 0$ and $d_j^s \neq 0$. Since $d_j^1, d_j^2, \ldots, d_j^s$ are conformal to d_j, either (i) $d_j^1 + \cdots + d_j^{s-1} > 0$ and $d_j^s > 0$ or (ii) $d_j^1 + \cdots + d_j^{s-1} < 0$ and $d_j^s < 0$. In case (i), we have $x_j + d_j^1 + \cdots + d_j^{s-1} < x_j + d_j$ and $x_j + d_j^s < x_j + d_j$, so the convexity of P_j [90, Lemma 2.12] implies

$$\begin{aligned} \frac{P_j(x_j + d_j^1 + \cdots + d_j^{s-1}) - P_j(x_j)}{d_j^1 + \cdots + d_j^{s-1}} &\leq \frac{P_j(x_j + d_j) - P_j(x_j)}{d_j}, \\ \frac{P_j(x_j + d_j^s) - P_j(x_j)}{d_j^s} &\leq \frac{P_j(x_j + d_j) - P_j(x_j)}{d_j}. \end{aligned}$$

Multiplying the above two inequalities by, respectively, $d_j^1 + \cdots + d_j^{s-1} > 0$ and $d_j^s > 0$ and summing, we have

$$P_j(x_j + d_j^1 + \cdots + d_j^{s-1}) - P_j(x_j) + P_j(x_j + d_j^s) - P_j(x_j) \leq P_j(x_j + d_j) - P_j(x_j). \quad (4.31)$$

In case (ii), we have $x_j + d_j^1 + \cdots + d_j^{s-1} > x_j + d_j$ and $x_j + d_j^s > x_j + d_j$, so the convexity of P_j implies

$$\begin{aligned} \frac{P_j(x_j + d_j^1 + \cdots + d_j^{s-1}) - P_j(x_j)}{d_j^1 + \cdots + d_j^{s-1}} &\geq \frac{P_j(x_j + d_j) - P_j(x_j)}{d_j}, \\ \frac{P_j(x_j + d_j^s) - P_j(x_j)}{d_j^s} &\geq \frac{P_j(x_j + d_j) - P_j(x_j)}{d_j}. \end{aligned}$$

Multiplying the above two inequalities by, respectively, $d_j^1 + \cdots + d_j^{s-1} < 0$ and $d_j^s < 0$ and summing, we again obtain (4.31). Since (4.30) holds for $r < s$, we also have

$$P_j(x_j + d_j^1 + \cdots + d_j^{s-1}) - P_j(x_j) \geq \sum_{t=1}^{s-1} \left(P_j(x_j + d_j^t) - P_j(x_j) \right).$$

Combining this with (4.31) proves that (4.30) holds for $r = s$. ∎

The assumption of P being separable is essential in Lemma 4.5. If we drop this assumption, then Lemma 4.5 is false. For example, take $P(x) = \|x\|$, $A = (1, 1, 1)$,

$b = 0$, $x = 0$, and $d = (1, 1, -2)^T$, let $d = d^1 + d^2 = (1, 0, -1)^T + (0, 1, -1)^T$, then $P(x + d) - P(x) = \sqrt{6} < 2\sqrt{2} = \sum_{t=1}^{2} \Big(P(x + d^t) - P(x)\Big)$.

By using Lemma 4.5 and generalizing the proof of Proposition 3.2, we obtain the following main result of this section.

Proposition 4.2 *For any $x \in X \cap \mathrm{dom} P$, $\ell \in \{\mathrm{rank}(A) + 1, \ldots, n\}$, and diagonal $D \succ 0$, there exists a nonempty $\mathcal{J} \subseteq \mathcal{N}$ satisfying $|\mathcal{J}| \leq \ell$ and*

$$q_D(x; \mathcal{J}) \leq \frac{1}{n - \ell + 1} q_D(x; \mathcal{N}). \tag{4.32}$$

Proof. Let $d = d_D(x; \mathcal{N})$. We divide our argument into three cases.

Case (i) $d = 0$: Then $q_D(x; \mathcal{N}) = 0$. Thus, for any nonempty $\mathcal{J} \subseteq \mathcal{N}$ with $|\mathcal{J}| \leq \ell$, we have from (4.5) and Lemma 4.1 with $H = D$ that $q_D(x; \mathcal{J}) \leq 0 = q_D(x; \mathcal{N})$, so (4.32) holds.

Case (ii) $d \neq 0$ and $|\mathrm{supp}(d)| \leq \ell$ (see Section 3.5 for the definition of *supp*(d)): Then $\mathcal{J} = \mathrm{supp}(d)$ satisfies $q_D(x; \mathcal{J}) = q_D(x; \mathcal{N})$ and hence (4.32), as well as $|\mathcal{J}| \leq \ell$.

Case (iii) $d \neq 0$ and $|\mathrm{supp}(d)| > \ell$: Since $d \in \mathrm{Null}(A)$, it has a conformal realization [87], [89, Section 10B], namely,

$$d = v^1 + \cdots + v^s,$$

for some $s \geq 1$ and some nonzero elementary vectors (see Section 3.5 for the definition of a elementary vector) $v^t \in \mathrm{Null}(A)$, $t = 1, \ldots, s$, conformal to d (see (3.33)). Then for some $\alpha > 0$, $\mathrm{supp}(d')$ is a proper subset of $\mathrm{supp}(d)$ and $d' \in \mathrm{Null}(A)$, where $d' = d - \alpha v^1$. (Note that αv^1 is an elementary vector of $\mathrm{Null}(A)$, so that $|\mathrm{supp}(\alpha v^1)| \leq \mathrm{rank}(A) + 1 \leq \ell$.) We repeat the above reduction step with d' in place of d. Since $|\mathrm{supp}(d')| \leq |\mathrm{supp}(d)| - 1$, after at most $|\mathrm{supp}(d)| - \ell$ reduction steps, we obtain

$$d = d^1 + \cdots + d^r, \tag{4.33}$$

for some $r \leq |\mathrm{supp}(d)| - \ell + 1$ and some nonzero $d^t \in \mathrm{Null}(A)$ conformal to d with $|\mathrm{supp}(d^t)| \leq \ell$, $t = 1, \ldots, r$. Since $|\mathrm{supp}(d)| \leq n$, we have $r \leq n - \ell + 1$.

Since $l - x \leq d \leq u - x$, (4.33) and d^t being conformal to d imply that $l - x \leq d^t \leq u - x$ for $t = 1, \ldots, r$. Since $Ad^t = 0$, this implies $x + d^t \in X$, $t = 1, \ldots, r$. Also (4.5) and (4.33) imply that

$$\begin{aligned}
q_D(x; \mathcal{N}) &= g^T d + \frac{1}{2} d^T D d + cP(x + d) - cP(x) \\
&= \sum_{t=1}^{r} g^T d^t + \frac{1}{2} \sum_{s=1}^{r} \sum_{t=1}^{r} (d^s)^T D d^t + cP\left(x + \sum_{t=1}^{r} d^t\right) - cP(x) \\
&\geq \sum_{t=1}^{r} g^T d^t + \frac{1}{2} \sum_{t=1}^{r} (d^t)^T D d^t + cP\left(x + \sum_{t=1}^{r} d^t\right) - cP(x) \\
&\geq \sum_{t=1}^{r} g^T d^t + \frac{1}{2} \sum_{t=1}^{r} (d^t)^T D d^t + \sum_{t=1}^{r} \Big(cP(x + d^t) - cP(x)\Big) \\
&\geq r \min_{t=1,\ldots,r} \left\{ g^T d^t + \frac{1}{2} (d^t)^T D d^t + cP(x + d^t) - cP(x) \right\},
\end{aligned}$$

where $g = \nabla f(x)$ and the first inequality uses (3.33) and $D \succ 0_n$ being diagonal, so that $(d^s)^T D d^t \geq 0$ for all s, t; the second inequality uses Lemma 4.5. Thus, if we let $\bar{t}$ be an index t attaining the above minimum and let $\mathcal{J} = \mathrm{supp}(d^{\bar{t}})$, then $|\mathcal{J}| \leq \ell$ and

$$\frac{1}{r} q_D(x; \mathcal{N}) \geq g^T d^{\bar{t}} + \frac{1}{2} (d^{\bar{t}})^T D d^{\bar{t}} + cP(x + d^{\bar{t}}) - cP(x) \geq q_D(x; \mathcal{J}),$$

where the second inequality uses $x + d^{\bar{t}} \in X$ and $d^{\bar{t}}_j = 0$ for $j \notin \mathcal{J}$. ∎

It can be seen from its proof that Proposition 4.2 still holds if the diagonal matrix D is only positive semidefinite, provided that $q_D(x; \mathcal{N}) > -\infty$ (such as when X is bounded).

The proof of Proposition 4.2 suggests, for any $\ell \in \{\mathrm{rank}(A)+1, \ldots, n\}$, an $O(n-\ell)$-step reduction procedure for finding a conformal realization (4.33) of $d = d_D(x; \mathcal{N})$ with $r \leq n - \ell + 1$ and a corresponding $\mathcal{J}$ satisfying $|\mathcal{J}| \leq \ell$ and (4.32). In the case of $m = 1$ and $\ell = 2$, we can find such a conformal realization in $O(n)$ operations. In the case of $m = 2$ and $\ell = 3$, $O(n \log n)$ operations are needed to find such a conformal realization. In general, the time complexity of finding such a conformal realization is $O(m^3(n - \ell)^2)$ operations; see Section 3.6 for more details.

In the case of $P \equiv 0$, the problem (1.13) is a linearly constrained smooth optimization. Then, for diagonal $D \succ 0$, $d_D(x;\mathcal{N})$ can be found in $O(n)$ operations; see Section 3.6 for more details.

If P in (1.1) is polyhedral such as $\|x\|_1$, then the corresponding subproblem (4.1) can be reformulated as a separable quadratic programming problem as described in [69], though the dimension can be more than double. Then, for diagonal $D \succ 0$, $d_D(x;\mathcal{N})$ can be found using an algorithm described by Megiddo and Tamir [69], which reportedly requires only $O(n)$ operations for each fixed m.

By combining the above observations, for the case where P is separable and polyhedral, we conclude that, for $m = 1$ and $\ell = 2$, an index subset $\mathcal{J}$ satisfying $|\mathcal{J}| \leq \ell$ and (4.32) can be found in $O(n)$ operations and, for $m \geq 1$ and $\ell \in \{\text{rank}(A)+1, \ldots, n\}$, such an index subset $\mathcal{J}$ can be found in $O(n^2)$ operations, where the constant in $O(\cdot)$ depends on m. It is an open question whether such a $\mathcal{J}$ can be found in $O(n)$ operations for a fixed $m \geq 2$.

In addition, if f is convex with Lipschitz continuous gradient, then, for $m = 1$ and $\ell = 2$, our overall complexity bound for achieving ϵ-optimality is given by $O\left(\frac{n^3 L b_{\max}^2}{\epsilon} + n^2 L \max\left\{0, \ln\left(\frac{e^0}{n b_{\max}}\right)\right\}\right)$ operations where $b_{\max} = \max_{1 \leq i \leq n}(u_i - l_i)$. Hence the number of operations required to be within ϵ of the optimal value grows cubic in the number n of variables. When specialized to the training of support vector machines, i.e., for the case where $m = 1$, $P \equiv 0$, and f is quadratic, our overall complexity bound reduces to $O\left(\frac{n^3 \Lambda b_{\max}^2}{\epsilon} + n^2 \Lambda \max\left\{0, \ln\left(\frac{e^0}{n b_{\max}}\right)\right\}\right)$ operations where Λ is the maximum norm of the 2×2 principal submatrices of $\nabla^2 f(x)$. For the large quadratic problem of the training of support vector machines, Hush and Scovel [42] proposed a decomposition method and proved that, for any $\epsilon > 0$, the overall complexity bound is $O(n^3 \ln n b_{\max}^2 (e^0 + n^2 \Lambda)/\epsilon)$ operations. This complexity bound was further improved by List and Simon [58] to problems with general linear constraints, where they showed that the overall complexity bound is $O\left(\frac{n^3 \Lambda b_{\max}^2}{\epsilon} + n^2 \max\left\{0, \ln\left(\frac{e^0}{n \Lambda b_{\max}}\right)\right\}\right)$ operations. Our complexity bound is as good as that of List and Simon [58] for the quadratic

problem of the training of support vector machines.

4.6 Conclusions and Extensions

We have proposed a block coordinate gradient descent method for linearly constrained nonsmooth minimization, have established its asymptotic linear convergence to a stationary point, and have given an upper bound on the number of iterations for the CGD method to achieve ϵ-optimality.

There are many directions for future research. For a diagonal D^k, we can find an index subset $\mathcal{J}^k$ satisfying (4.6) in polynomial time if P is separable and polyhedral. Can we use a nondiagonal D^k and still efficiently find a $\mathcal{J}^k$ satisfying (4.6)? If $P(x) = \|x\|$, then the Lemma 4.5 is not satisfied in general. Can a result similar to Proposition 4.2 be proved when $P(x) = \|x\|$ or, more generally, when $P(x)$ is nonseparable. Can we efficiently find an index subset if P is nonseparable?

The problem (1.13) can be generalized to the following problem:

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & f(x) + cP(x) \\ \text{s.t.} & f_1(x) = 0, \ldots, f_m(x) = 0, \end{array}$$

where $c > 0$, $P : \mathbb{R}^n \to (-\infty, \infty]$ is a (separable) proper, convex, lsc function, and $f_1(x), \ldots, f_m(x)$ are twice continuously differentiable functions. Can the CGD method be extended to solve this more general problem?

Chapter 5

A (BLOCK) COORDINATE GRADIENT DESCENT METHOD FOR BI-LEVEL OPTIMIZATION

In this chapter, we study ways to dynamically adjust c in (1.1) towards 0 in the CGD method so as to solve (1.15). First, we briefly review the CGD method. Then we describe an algorithm for solving (1.15) and show that any cluster point of the generated iterates is a solution of the bi-level problem (1.15) if the smooth function f is convex, the convex function P is proper, level-bounded, lsc, and the set of stationary points of f over the domain of P is nonempty. This chapter is based on the paper [105] co-authored with P. Tseng.

5.1 (Block) Coordinate Gradient Descent Method

In this section, we briefly review the CGD method for solving the regularized problem (see Section 2.1 for more details):

$$\min_x \quad F_c(x) \stackrel{\text{def}}{=} f(x) + cP(x). \tag{5.1}$$

In the CGD method, we use $\nabla f(x)$ to build a quadratic approximation of f at x and apply coordinate descent to generate an improving direction d at x. More precisely, we choose a nonempty index subset $\mathcal{J} \subseteq \mathcal{N}$ and a symmetric matrix $H \succ 0_n$ (approximating the Hessian $\nabla^2 f(x)$), and move x along the direction $d = d_H(x; c; \mathcal{J})$, where

$$d_H(x; c; \mathcal{J}) \stackrel{\text{def}}{=} \operatorname*{arg\,min}_{d_j=0 \ \forall j \notin \mathcal{J}} \left\{ \nabla f(x)^T d + \frac{1}{2} d^T H d + cP(x+d) - cP(x) \right\}. \tag{5.2}$$

Notice that $d_H(x; c; \mathcal{J})$ depends on H only through $H_{\mathcal{J}\mathcal{J}}$.

For convergence, the index subset $\mathcal{J}$ must be chosen judiciously. We propose the following three rules for choosing $\mathcal{J}$:

- *Gauss-Seidel* : $\mathcal{J}$ cycles through $\{1\}, \{2\}, ..., \{n\}$ or, more generally, every T consecutive such $\mathcal{J}$ collectively covers $1, 2, ..., n$, where $T \geq 1$ [13, 39, 63, 78, 102].

- *Gauss-Southwell-r* : We choose $\mathcal{J}$ to satisfy

$$\|d_D(x; c; \mathcal{J})\|_\infty \geq \upsilon \|d_D(x; c; \mathcal{N})\|_\infty,$$

  where $0 < \upsilon \leq 1$ and $D \succ 0_n$ is diagonal.

- *Gauss-Southwell-q* : For any $x \in \mathrm{dom}P$, nonempty $\mathcal{J} \subseteq \mathcal{N}$, and $H \succ 0$, define $q_H(x; c; \mathcal{J})$ to be the optimal objective value of (5.2). Thus $q_H(x; c; \mathcal{J})$ estimates the descent in F_c from x to $x + d_H(x; c; \mathcal{J})$. We choose $\mathcal{J}$ to satisfy

$$q_D(x; c; \mathcal{J}) \leq \upsilon \; q_D(x; c; \mathcal{N}),$$

  where $0 < \upsilon \leq 1$, $D \succ 0_n$ is diagonal.

The global convergence of the CGD method using the above three rules are proved in Section 2.3.

Formally, we say that $x \in \mathbb{R}^n$ is *a stationary point* of F_c if $x \in \mathrm{dom}F_c$ and $F_c'(x; d) \geq 0$ for all $d \in \mathbb{R}^n$. The following lemma gives an alternative characterization of stationarity.

Lemma 5.1 *For any $H \succ 0_n$, an $x \in \mathrm{dom}P$ is a stationary point of F_c if and only if $d_H(x; c; \mathcal{N}) = 0$.*

Proof. See the proof of Lemma 2.1. ■

Thus, $\|d_H(x; c; \mathcal{N})\|$ acts as a scaled "residual" function (with scaling matrix H), measuring how close x comes to being stationary for F_c. Hence $\|d_H(x; c; \mathcal{N})\|$ can be used for a measure of current solution accuracy; see Section 5.2.

5.2 CGD-Homotopy Method and Convergence Analysis

In this section, we describe an algorithm, which uses the CGD method to generate an approximate solution of (5.1) for fixed c, for solving (1.15) and establish that any cluster point of the generated iterates is a solution of (1.15).

For any $x \in \mathrm{dom}P$, let

$$R_c(x) \stackrel{\mathrm{def}}{=} d_I(x; c; \mathcal{N}). \tag{5.3}$$

Assume the set of stationary points $\bar{X}$ of (5.1) is nonempty. By the continuity of ∇f and Lemma 5.1 with $H = I$ and $c = c^k$, for $\epsilon^k > 0$, there exists an approximate solution x that satisfies

$$\|R_{c^k}(x)\| \leq \epsilon^k, \tag{5.4}$$

$$-(x + \nabla f(x))^T R_{c^k}(x) \leq \epsilon^k. \tag{5.5}$$

Our method for solving (1.15) uses similar idea as in [93] for a primal-dual interior-point method. At each iteration k ($k = 0, 1, 2, \ldots$), a regularization parameter $c^k > 0$ and an accuracy tolerance ϵ^k are chosen, and the CGD method is applied to (5.1) with $c = c^k$ until it finds an approximate solution x^k satisfying the conditions (5.4) and (5.5). Since the idea of decreasing c is reminiscent of homotopy methods for equation solving, we call this the CGD-homotopy method.

CGD-Homotopy Method:

Choose $x^0 \in \mathrm{dom}P$, $c^0 > 0$, $\epsilon^0 > 0$. For $k = 1, 2, \ldots$, generate x^k from x^{k-1} according to the outer iteration:

1. Choose $c^k > 0$ and $\epsilon^k > 0$.

2. Compute an $x^k \in \mathrm{dom}P$ satisfying (5.4) and (5.5) by applying the CGD method to (5.1) with $c = c^k$ and an initial point $x = x^{k-1}$.

The following lemma shows that the bi-level problem (1.15) has a solution and the optimal objective value is finite if P is a level-bounded function and $S_f \neq \emptyset$.

Lemma 5.2 *Suppose P is level-bounded and $S_f \neq \emptyset$. Then the minimum of P over S_f is finite and attained on a nonempty compact subset of S_f.*

Proof. Let $\tilde{P}(x) = P(x) + \delta_{S_f}(x)$, where $\delta_{S_f}(x)$ is the indicator function of the set S_f, then $\tilde{P}$ is proper because $S_f \subseteq \mathrm{dom} P$ and $S_f \neq \emptyset$, and it's lsc by [90, Theorem 1.6] because its level sets of the form $S_f \cap \{x \mid P(x) \leq \xi\}$ for $\xi \leq \infty$ are closed (by virtue of the closedness of S_f and the lower semicontinuity of P). Since P is level-bounded, sets of the form $S_f \cap \{x \mid P(x) \leq \xi\}$ for $\xi \leq \infty$ are bounded. Hence $\tilde{P}$ is level-bounded. By [90, Theorem 1.9], the minimum of $\tilde{P}$ is finite and the $\arg\min \tilde{P}$ is nonempty and compact. Therefore the minimum of P over S_f is finite and attained on a nonempty compact subset of S_f. ∎

The following theorem shows that, by letting $c^k \to 0$ and $\epsilon^k \to 0$ at suitable rates in the CGD-homotopy method, every cluster point of the approximate solutions $\{x^k\}$ solves (1.15).

Theorem 5.1 *Assume that f is convex, P is level-bounded and $S_f \neq \emptyset$. If we choose c^k and ϵ^k to tend to zero such that*

$$\lim_{k\to\infty} \frac{\epsilon^k}{c^k} = 0, \tag{5.6}$$

then every cluster point of $\{x^k\}$ is a solution of (1.15).

Proof. Let $x^* \in \arg\min_{x\in S_f} P(x)$. At iteration k, we choose approximate solution x^k satisfying (5.4) and (5.5).

By Fermat's rule [90, Theorem 10.1] and (5.3),

$$R_{c^k}(x^k) \in \arg\min_d \ (g^k + R_{c^k}(x^k))^T d + c^k P(x^k + d) - c^k P(x^k),$$

where $g^k = \nabla f(x^k)$. Hence

$$\begin{aligned} & (g^k + R_{c^k}(x^k))^T R_{c^k}(x^k) + c^k P(x^k + R_{c^k}(x^k)) - c^k P(x^k) \\ \leq \ & (g^k + R_{c^k}(x^k))^T (x^* - x^k) + c^k P(x^*) - c^k P(x^k). \end{aligned}$$

Adding $f(x^k)$ and $c^k P(x^k)$ on both sides of the above inequality and rearranging terms yield

$$\begin{aligned} & f(x^k) + c^k P(x^k + R_{c^k}(x^k)) + R_{c^k}(x^k)^T R_{c^k}(x^k) \\ \leq \; & f(x^k) + (g^k + R_{c^k}(x^k))^T (x^* - x^k) + c^k P(x^*) - (g^k)^T R_{c^k}(x^k). \end{aligned} \tag{5.7}$$

Since f is convex, $f(x^k) + (g^k)^T(x^* - x^k) \leq f(x^*)$. This together with (5.4), (5.5), (5.7), and $R_{c^k}(x^k)^T R_{c^k}(x^k) \geq 0$ implies that

$$\begin{aligned} & f(x^k) + c^k P(x^k + R_{c^k}(x^k)) \\ \leq \; & f(x^*) + R_{c^k}(x^k)^T (x^* - x^k) + c^k P(x^*) - (g^k)^T R_{c^k}(x^k) \\ \leq \; & f(x^*) + \|R_{c^k}(x^k)\| \|x^*\| - (x^k + g^k)^T R_{c^k}(x^k) + c^k P(x^*) \\ \leq \; & f(x^*) + \epsilon^k \|x^*\| + \epsilon^k + c^k P(x^*). \end{aligned} \tag{5.8}$$

Since $x^* \in S_f$ and $x^k \in \mathrm{dom} P$, $f(x^*) \leq f(x^k)$. This together with (5.8) implies

$$c^k P(x^k + R_{c^k}(x^k)) \leq \epsilon^k \|x^*\| + \epsilon^k + c^k P(x^*).$$

Dividing both sides of the above inequality by c^k yields

$$P(x^k + R_{c^k}(x^k)) \leq \frac{\epsilon^k}{c^k} \|x^*\| + \frac{\epsilon^k}{c^k} + P(x^*). \tag{5.9}$$

Since P is level-bounded, by (5.6), $\{x^k + R_{c^k}(x^k)\}$ is bounded as $k \to \infty$. This together with $R_{c^k}(x^k) \to 0$ (since x^k satisfies (5.4)) implies that $\{x^k\}$ has cluster points. By (5.8), $c^k \to 0$ and $\epsilon^k \to 0$, we see that any cluster point $\bar{x}$ of $\{x^k\}$ satisfies $f(\bar{x}) \leq f(x^*)$. Thus $\bar{x} \in S_f$. Moreover, (5.6) and (5.9) and the lsc property of P imply $P(\bar{x}) \leq P(x^*)$. Thus $\bar{x}$ solves (1.15). ■

5.3 Conclusions and Extensions

We have proposed a regularization strategy for solving the bi-level problem with the regularized problem solved by the CGD method and have established its global

convergence to a solution under convexity assumption on f and level-boundedness assumption on P in addition to be proper, convex, and lsc.

Can any one of the assumptions on P in Theorem 5.1 (i.e., proper, convex, lsc, level-bound) be dropped? The global convergence for the CGD method is still satisfied when f is nonconvex. Can the regularization strategy be extended to handle a nonconvex function, i.e., f is nonconvex?

BIBLIOGRAPHY

[1] Antoniadis, A. and Fan, J., Regularization of wavelet approximations, J. Amer. Statist. Assoc. 96 (2001), 939–967.

[2] Auslender, A., Minimisation de fonctions localement lipschitziennes: applications à la programmation mi-convexe, mi-différentiable, in O. L. Mangasarian, R. R. Meyer and S. M. Robinson, editors, Nonlinear Programming, 3, Academic Press, New York, 1978, 429–460.

[3] Barr, R. O. and Gilbert. E. G., Some efficient algorithms for a class of abstract optimization problems arising in optimal control, IEEE Trans. Auto. Control 14 (1969), 640–652.

[4] Berman, P., Kovoor, N., and Pardalos, P. M., Algorithms for the least distance problem, in Complexity in Numerical Optimization, P. M. Pardalos, ed., World Scientific, Singapore, 1993, 33–56.

[5] Bertsekas, D. P., Constrained Optimization and Lagrange Multiplier Methods, Academic Press, New York, 1982.

[6] Bertsekas, D. P., Nonlinear Programming, 2nd edition, Athena Scientific, Belmont, 1999.

[7] Bradley, P. S., Fayyad, U. M., and Mangasarian, O. L., Mathematical programming for data mining: formulations and challenges, INFORMS J. Comput. 11 (1999), 217–238.

[8] Brucker, P., An $O(n)$ algorithm for quadratic knapsack problems, Oper. Res. Letters 3 (1984), 163–166.

[9] Burke, J. V., Descent methods for composite nondifferentiable optimization problems, Math. Prog. 33 (1985), 260–279.

[10] Cands, E. J., Romberg, J., and Tao, T., Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Info. Theory 52 (2006), 489–509.

[11] Chang, C.-C., Hsu, C.-W., and Lin, C.-J., The analysis of decomposition methods for support vector machines, IEEE Trans. Neural Networks 11 (2000), 1003–1008.

[12] Chang, C.-C. and Lin, C.-J., LIBSVM: a library for support vector machines, 2001, available from `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[13] Censor, Y. and Zenios, S. A., Parallel Optimization: Theory, Algorithms, and Applications, Oxford Univ. Press, New York, 1997.

[14] Chen, S., Donoho, D., and Saunders, M., Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. 20 (1999), 33–61.

[15] Chen, P.-H., Fan, R.-E., and Lin, C.-J., A study on SMO-type decomposition methods for support vector machines, IEEE Trans. Neural Networks 17 (2006), 893–908.

[16] Coleman, T. F. and Li, Y., An interior trust region approach for nonlinear minimization subject to bounds, SIAM J. Optim. 6 (1996), 418–445.

[17] Conn, A. R., Gould, N. I. M., and Toint, Ph. L., Trust-Region Methods, SIAM, Philadelphia, 2000.

[18] Correa, J. R., Schulz, A. S., and Stier Moses, N. E., Selfish routing in capacitated networks, Math. Oper. Res. 29 (2004), 961–976.

[19] Cristianini, N. and Shawe-Taylor, J., An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, 2000.

[20] Donoho, D. L. and Johnstone, I. M., Ideal spatial adaptation by wavelet shrinkage, Biometrika 81 (1994), 425–455.

[21] Donoho, D. L. and Johnstone, I. M., Adapting to unknown smoothness via wavelet shrinkage, J. Amer. Statist. Assoc. 90 (1995), 1200–1224.

[22] Donoho, D. and Tanner, J., Sparse nonnegative solutions of underdetermined linear equations by linear programming, Proc. Nat. Acad. Sci. USA 102 (2005), 9446–9451.

[23] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik, V., Support vector regression machines, in M. C. Mozer, M. I. Jordan, and T. Petsche, editors, Advances in Neural Information Processing Systems 9, MIT Press, Cambridge, 1997.

[24] Facchinei, F., Fischer, A., and Kanzow, C., On the accurate identification of active constraints, SIAM J. Optim. 9 (1998), 14–32.

[25] Facchinei, F. and Pang, J.-S., Finite-Dimensional Variational Inequalities and Complementarity Problems, Vols. I and II, Springer-Verlag, New York, 2003.

[26] Fan, R.-E., Chen, P.-H., and Lin, C.-J., Working set selection using second order information for training support vector machines, J. Mach. Learn. Res. 6 (2005), 1889–1918.

[27] Ferris, M. C. and Mangasarian, O. L., Parallel variable distribution, SIAM J. Optim. 4 (1994), 815–832.

[28] Ferris, M. C. and Munson, T. S., Interior-point methods for massive support vector machines, SIAM J. Optim. 13 (2003), 783–804.

[29] Ferris, M. C. and Munson, T. S., Semismooth support vector machines, Math. Prog. 101 (2004), 185–204.

[30] Fine, S. and Scheinberg, K., Efficient SVM training using low-rank kernel representations, J. Mach. Learn. Res. 2 (2001), 243–264.

[31] Fletcher, R., A model algorithm for composite nondifferentiable optimization problems, Math. Prog. Study 17 (1982), 67–76.

[32] Fletcher, R., Practical Methods of Optimization, 2nd edition, John Wiley & Sons, Chichester, 1987.

[33] Fletcher, R., An overview of unconstrained optimization, in Algorithms for Continuous Optimization, edited by E. Spedicato, Kluwer Academic, Dordrecht, 1994, 109–143.

[34] Fukushima, M., A successive quadratic programming method for a class of constrained nonsmooth optimization problems, Math. Prog. 49 (1990/91), 231–251.

[35] Fukushima, M., Parallel variable transformation in unconstrained optimization, SIAM J. Optim. 8 (1998), 658–672.

[36] Fukushima, M. and Mine, H., A generalized proximal point algorithm for certain non-convex minimization problems, Int. J. Systems Sci. 12 (1981), 989–1000.

[37] Glasmachers, T. and Igel, C., Maximum-gain working set selection for SVMs, J. Mach. Learn. Res. 7 (2006), 1437–1466.

[38] Gould, N. I. M., Orban, D., and Toint, Ph. L., CUTEr, a constrained and unconstrained testing environment, revisited, ACM Trans. Math. Software 29 (2003), 373–394.

[39] Grippo, L. and Sciandrone, M., On the convergence of the block nonlinear Gauss-Seidel method under convex constraints, Oper. Res. Letters 26 (2000), 127–136.

[40] Han, S.-P., A successive projection method, Math. Prog. 40 (1988), 1–14.

[41] Han, S.-P., A decomposition method and its application to convex programming, Math. Oper. Res. 14 (1989), 237–248.

[42] Hush, D. and Scovel, C., Polynomial-time decomposition algorithms for support vector machines, Machine Learning 51 (2003), 51–71.

[43] Hush, D., Kelly, P., Scovel, C., and Steinwart, I., QP algorithms with guaranteed accuracy and run time for support vector machines, J. Mach. Learn. Res. 7 (2006), 733–769.

[44] Iusem, A. N., Pennanen, T., and Svaiter, B. F., Inexact variants of the proximal point algorithm without monotonicity, SIAM J. Optim. 13 (2003), 1080–1097.

[45] Joachims, T., Making large-scale SVM learning practical, in Advances in Kernel Methods - Support Vector Learning, B. Schólkopf, C. J. Burges, and A. J. Smola, eds., MIT Press, Cambridge, MA, 1998.

[46] Kao, C., Lee, L.-F., and Pitt, M. M., Simulated maximum likelihood estimation of the linear expenditure system with binding non-negativity constraints, Ann. Econ. Fin. 2 (2001), 203–223.

[47] Keerthi, S. S. and Gilbert, E. G., Convergence of a generalized SMO algorithm for SVM classifier design, Machine Learning 46 (2002), 351–360.

[48] Keerthi, S. S. and Ong, C. J., On the role of the threshold parameter in SVM training algorithm, Technical Report CD-00-09, Department of Mathematical and Production Engineering, National University of Singapore, Singapore, 2000.

[49] Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. K., Improvements to Platt's SMO algorithm for SVM classifier design, Neural Comput. 13 (2001), 637–649.

[50] Kelley, C. T., Iterative Methods for Optimization, SIAM, Philadelphia, 1999.

[51] Kiwiel, K. C., A method for minimizing the sum of a convex function and a continuously differentiable function, J. Optim. Theory Appl. 48 (1986), 437–449.

[52] Kiwiel, K. C., On linear time algorithms for the continuous quadratic knapsack problem, report, Systems Research Institute, Warsaw, Poland, 2006; to appear in J. Optim. Theory Appl.

[53] Lin, C.-J., On the convergence of the decomposition method for support vector machines, IEEE Trans. Neural Networks 12 (2001), 1288–1298.

[54] Lin, C.-J., Linear convergence of a decomposition method for support vector machines, technical report, Department of Computer Science and Information Engineering, Taiwan University, Taipei, Taiwan, 2001.

[55] Lin, C.-J., Asymptotic convergence of an SMO algorithm without any assumptions, IEEE Trans. Neural Networks 13 (2002), 248–250.

[56] Lin C.-J., Lucidi S., Palagi L., Risi A., and Sciandrone M., A decomposition algorithm model for singly linearly constrained problems subject to lower and upper bounds, technical report, DIS-Università di Roma "La Sapienza", Rome, January 2007; submitted to J. Optim. Theory Appl.

[57] List, N. and Simon, H. U., A general convergence theorem for the decomposition method, in Proceedings of the 17th Annual Conference on Learning Theory, 2004, 363–377.

[58] List, N. and Simon, H. U., General polynomial time decomposition algorithms, in Lecture Notes in Computer Science Volume 3559/2005, Springer, Berlin, 2005, 308–322.

[59] Lucidi, S., Palagi, L., Risi, A., and Sciandrone, M., On the convergence of hybrid decomposition methods for SVM training, technical report, DIS-Università di Roma "La Sapienza", Rome, July 2006; submitted to IEEE Trans. Neural Networks.

[60] Luo, Z.-Q. and Tseng, P., Error bounds and the convergence analysis of matrix splitting algorithms for the affine variational inequality problem, SIAM J. Optim. 2 (1992), 43–54.

[61] Luo, Z.-Q. and Tseng, P., On the linear convergence of descent methods for convex essentially smooth minimization, SIAM J. Control Optim. 30 (1992), 408–425.

[62] Luo, Z.-Q. and Tseng, P., On the convergence rate of dual ascent methods for linearly constrained convex minimization, Math. Oper. Res. 18 (1993), 846–867.

[63] Luo, Z.-Q. and Tseng, P., Error bounds and convergence analysis of feasible descent methods: a general approach, Ann. Oper. Res. 46 (1993), 157–178.

[64] Mangasarian, O. L, Sparsity-preserving SOR algorithms for separable quadratic and linear programming, Comput. Oper. Res. 11 (1984), 105–112.

[65] Mangasarian, O. L., Parallel gradient distribution in unconstrained optimization, SIAM J. Control Optim. 33 (1995), 1916–1925.

[66] Mangasarian, O. L. and De Leone, R., Parallel gradient projection successive overrelaxation for symmetric linear complementarity problems and linear programs, Ann. Oper. Res. 14 (1988), 41–59.

[67] Mangasarian, O. L. and Musicant, D. R., Successive overrelaxation for support vector machines, IEEE Trans. Neural Networks 10 (1999), 1032–1037.

[68] Mangasarian, O. L. and Musicant, D. R., Large scale kernel regression via linear programming, Machine Learning 46 (2002), 255–269.

[69] Megiddo, N. and Tamir, A., Linear time algorithms for some separable quadratic programming problems, Oper. Res. Letters 13 (1993), 203–211.

[70] Meier, L., van de Geer, S., and Bühlmann, P., The group Lasso for logistic regression. Report, Seminar für Statistik, ETH Zürich, Zürich, March 2006.

[71] Meyer, R. R., Multipoint methods for separable nonlinear networks, Math. Prog. Study 22 (1985), 185–205.

[72] Mine, H. and Fukushima, M., A minimization method for the sum of a convex function and a continuously differentiable function, J. Optim. Theory Appl. 33 (1981), 9–23.

[73] Moré, J. J., Garbow, B. S., and Hillstrom, K. E., Testing unconstrained optimization software, ACM Trans. Math. Software 7 (1981), 17–41.

[74] Moré, J. J. and Toraldo, G., On the solution of large quadratic programming problems with bound constraints, SIAM J. Optim. 1 (1991), 93–113.

[75] Murtagh, B. A. and Saunders, M. A., MINOS 5.5 user's guide, Report SOL 83-20R, Department of Operations Research, Stanford University, Stanford (Revised July 1998).

[76] Nocedal, J., Updating quasi-Newton matrices with limited storage, Math. Comp. 35 (1980), 773–782.

[77] Nocedal, J. and Wright S. J., Numerical Optimization, Springer-Verlag, New York, 1999.

[78] Ortega, J. M. and Rheinboldt, W. C., Iterative Solution of Nonlinear Equations in Several Variables, reprinted by SIAM, Philadelphia, 2000.

[79] Osuna, E., Freund, R., and Girosi, F., Improved training algorithm for support vector machines, Proc. IEEE NNSP '97, 1997.

[80] Palagi, L. and Sciandrone, M., On the convergence of a modified version of $\text{SVM}^{\text{light}}$ algorithm, Optim. Methods Softw. 20 (2005), 317–334.

[81] Pennanen, T., Local convergence of the proximal point algorithm and multiplier methods without monotonicity, Math. Oper. Res. 27 (2002), 170–191.

[82] Platt, J., Sequential minimal optimization: A fast algorithm for training support vector machines, in Advances in Kernel Methods-Support Vector Learning, B. Schölkopf, C. J. C. Burges, and A. J. Smola, eds. MIT Press, Cambridge, MA, 1999, 185–208.

[83] Powell, M. J. D., On search directions for minimization algorithms, Math. Prog. 4 (1973), 193–201.

[84] Robinson, S. M., Some continuity properties of polyhedral multifunctions, Math. Prog. Study 14 (1981), 206–214.

[85] Robinson, S. M., Linear convergence of ϵ-subgradient descent methods for a class of convex functions, Math. Prog. 86 (1999), 41–50.

[86] Robinson, S. M., Calmness and Lipschitz continuity for multifunctions, Report, Department of Industrial Engineering, University of Wisconsin, Madison, 2006.

[87] Rockafellar, R. T., The elementary vectors of a subspace of R^N, in Combinatorial Mathematics and its Applications, Proc. of the Chapel Hill Conference 1967, R. C. Bose and T. A. Dowling, editors, Univ. North Carolina Press, Chapel Hill, NC, 1969, 104–127.

[88] Rockafellar, R. T., Convex Analysis, Princeton University Press, Princeton, 1970.

[89] Rockafellar, R. T., Network Flows and Monotropic Optimization, Wiley-Interscience, New York, 1984; republished by Athena Scientific, Belmont, MA, 1998.

[90] Rockafellar, R. T. and Wets R. J.-B., Variational Analysis, Springer-Verlag, New York, 1998

[91] Sardy, S., Bruce, A., and Tseng, P., Block coordinate relaxation methods for nonparametric wavelet denoising, J. Comput. Graph. Stat. 9 (2000), 361–379.

[92] Sardy, S., Bruce, A., and Tseng, P., Robust wavelet denoising, IEEE Trans. Signal Proc. 49 (2001), 1146–1152.

[93] Sardy, S. and Tseng, P., AMlet, RAMlet, and GAMlet: automatic nonlinear fitting of additive models, robust and generalized, with wavelets, J. Comput. Graph. Statist. 13 (2004), 283–309.

[94] Sardy, S. and Tseng, P., On the statistical analysis of smoothing by maximizing dirty Markov random field posterior distributions, J. Amer. Statist. Assoc. 99 (2004), 191–204.

[95] Saunders, C., Stitson, M. O., Weston, J., Bottou, L., Schölkopf., B. and Smola, A., Support vector machine – reference manual, Report CSD-TR-98-03, Department of Computer Science, Royal Holloway, University of London, Egham, UK, 1998.

[96] Scheinberg, K., An efficient implementation of an active set method for SVM, J. Mach. Learn. Res. 7 (2006), 2237–2257.

[97] Schölkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L., New support vector algorithms, Neural Comput. 12 (2000), 1207–1245.

[98] Simon, H. U., On the complexity of working set selection, Proceedings of the 15th International Conference on Algorithmic Learning Theory, 2004, 324–337.

[99] Spingarn, J. E., Submonotone mappings and the proximal point algorithm, Numer. Funct. Anal. Optim. 4 (1981/82), 123–150.

[100] Tseng, P., On the rate of convergence of a partially asynchronous gradient projection algorithm, SIAM J. Optim. 1 (1991), 603–619.

[101] Tseng, P., Dual coordinate ascent methods for non-strictly convex minimization, Math. Prog. 59 (1993), 231–247.

[102] Tseng, P., Convergence of block coordinate descent method for nondifferentiable minimization, J. Optim. Theory Appl. 109 (2001), 473–492.

[103] Tseng, P. and Yun S., A coordinate gradient descent method for nonsmooth separable minimization, report, Department of Mathematics, University of Washington, Seattle, June 2006; submitted to Math. Program. B.

[104] Tseng, P. and Yun S., A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training, report, Department of Mathematics, University of Washington, Seattle, March 2007, submitted to Comput. Optim. Appl.

[105] Tseng, P. and Yun S., A coordinate gradient descent method for constrained nonsmooth optimization and bi-level optimization, June 2007.

[106] Vapnik, V., Estimation of Dependences Based on Empirical Data, Springer-Verlag, New York, 1982.

[107] Vapnik, V., Golowich, S. E., and Smola, A., Support vector method for function approximation, regression estimation, and signal processing, in M. C. Mozer, M. I. Jordan, and T. Petsche, editors, Advances in Neural Information Processing Systems 9, MIT Press, Cambridge, 1997.

[108] Yuan, L. and Lin, Y., Model selection and estimation in regression with grouped variables. J. Royal Stat. Soc. 68 (2006), 49–67.

[109] Zhu, C., Byrd, R. H., and Nocedal, J., L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization, ACM Trans. Math. Software 23 (1997), 550–560.

VITA

Sangwoon Yun was born in Seoul, Korea, on September 4th of 1971. He received a B.S. and a M.S. in Mathematics from Yonsei University in 1995 and 1999, respectively. He was married to Jinhee Hong on June 23rd of 1994 and blessed his first son, Paul, in September of 2005. He entered the University of Washington in the Autumn of 2001, and expects to obtain a Ph.D. in Mathematics in August of 2007.